AD-A242 921

DTIC

**U.S. Army Research Institute
for the Behavioral and Social Sciences**

Research Report 1597

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Final Report on Project A

John P. Campbell and Lola M. Zook, Editors
Human Resources Research Organization

91-16420

August 1991

91 1125 022

U.S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction
of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

MICHAEL D. SHALER
COL, AR
Commanding

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Michael G. Rumsey
Frances C. Grafton

NOTICES

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --

2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	

4. PERFORMING ORGANIZATION REPORT NUMBER(S) FR-PRD-90-06	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Report 1597

6a. NAME OF PERFORMING ORGANIZATION Human Resources Research Organization (HumRRO)	6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute

6c. ADDRESS (City, State, and ZIP Code) 1100 South Washington Street Alexandria, VA 22314-4499	7b. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600

8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences	8b. OFFICE SYMBOL (If applicable) PERI-R	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-82-C-0531

8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO. 63007A	PROJECT NO. 792	TASK NO. 232	WORK UNIT ACCESSION NO. C1

11. TITLE (Include Security Classification)
Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Final Report on Project A

12. PERSONAL AUTHOR(S)
Campbell, John P.; and Zook, Lola M., editors

13a. TYPE OF REPORT Final	13b. TIME COVERED FROM 82/10 TO 90/03	14. DATE OF REPORT (Year, Month, Day) 1991, August	15. PAGE COUNT

16. SUPPLEMENTARY NOTATION
Prepared under Project A: Improving the selection, classification, and utilization of Army enlisted personnel (HumRRO, AIR, Personnel Decisions Research Institute, ARI).

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Army-wide measures First-tour performance
			Criterion measures Longitudinal validation
			Data collection Personnel classification (Continued)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)
This report describes the research conducted during Project A, a research project that represented the first phase of the Army's long-term program to develop a complete personnel system for selecting and classifying entry-level enlisted personnel. The goal of Project A was to increase effectiveness in matching first-tour enlisted manpower requirements with available personnel resources through validation of existing selection and classification tests and development of new and improved tests that will predict measures of job performance, including aspects of second-tour performance. The project, which began in 1982, was under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). The research included three main phases (a) evaluation of FY81/82 accessions' scores on the Armed Services Vocational Aptitude Battery (ASVAB) and subsequent Army performance; (b) selection of 19 Military Occupational Specialties (MOS) as a representative sample of the Army's 250+ entry-level MOS, development and field test of a preliminary battery of predictor-type tests and of a comprehensive set of job performance (Continued)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Michael G. Rumsey	22b. TELEPHONE (Include Area Code) (703) 274-8275	22c. OFFICE SYMBOL PERI-RS

DD Form 1473, JUN 86   Previous editions are obsolete.   SECURITY CLASSIFICATION OF THIS PAGE

ARI Research Report 1597

18. SUBJECT TERMS (Continued):

    Personnel selection
    Predictor measures
    Project A
    Rating scales
    Second-tour performance

19. ABSTRACT (Continued):

measures, and administration and evaluation of these instruments with FY83/84
Army accessions in the Concurrent Validation stage; and (c) administration of
the revised predictor battery to 49,000 accessions from 21 MOS during FY86/87
and subsequent administration to 11,000 soldiers from this group during their
first tour, in the Longitudinal Validation stage. Development and administra-
tion of second-tour measures was also begun under this project.

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Final Report on Project A

John P. Campbell and Lola M. Zook, Editors
Human Resources Research Organization

Selection and Classification Technical Area
Michael G. Rumsey, Chief

Manpower and Personnel Research Laboratory
Zita M. Simutis, Director

iii

DEDICATION

This final report is dedicated to the memory of Newell Kent Eaton, Ph.D. His name will always be linked to Project A. Much of what was accomplished was due to his vision, his vitality, and his dedication.

This document is a description of the project that represents the first phase of the Army's long-term research effort to improve the selection, classification, and utilization of Army enlisted personnel. The thrust for the project came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB--the current U.S. military selection/classification test battery) and other selection variables as predictors of training success and job performance.

The portion of the effort described herein was devoted to the development and validation of Army Selection and Classification measures, referred to as "Project A." Project A was conducted under contract by the Selection and Classification Technical Area (SCTA) of the Manpower and Personnel Research Laboratory (MPRL) at the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). The research supports the MPRL and SCTA mission to improve the Army's capability to select and classify its applicants for enlistment or reenlistment by ensuring that fair and valid measures are developed for evaluating applicant potential based on expected job performance and utility to the Army.

Project A was authorized through a letter, Deputy Chief of Staff for Operations, "Army Research Project to Validate the Predictive Value of the Armed Services Vocational Aptitude Battery," effective 19 November 1980 and a Memorandum, Assistant Secretary of Defense (MRA&L), "Enlistment Standards," effective 11 September 1980.

To ensure that Project A research achieved its full scientific potential and would be useful to the Army, an advisory group comprised of Army general officers, interservice scientists, and experts in personnel measurement, selection, and classification was established. Members of the expert component provided guidance on technical aspects of the research, while general officer and interservice components oversaw the entire research effort, provided military judgment, provided periodic reviews of the project's progress, results, and plans, and coordinated within their comman s. Members of the General Officers' Advisory Group varied during the 7-year period covered by this report. Throughout the course of the project, this group was briefed on the plans and results of the various research phases and provided continuing military guidance. Members of Project A's Scientific Advisory Group guided the technical quality of the research. During the period covered by this report members included Drs. Philip Bobko, Thomas Cook, Milton Hakel (Chair), Lloyd Humphreys, Lawrence Johnson, Robert Linn, Mary Tenopyr, and Jay Uhlaner. This group was briefed throughout the project on the technical concepts, plans, and implementation results and provided advice on the further development of classification and assignment principles and procedures.

This final report on Project A summarizes the development and evaluation work done during the three main phases of the research: (a) analysis of file data on an FY81/82 accession sample to compare their ASVAB scores and their subsequent Army performance; (b) selection of a representative sample of entry-level MOS, and development and testing of predictor and job performance measures with a sample of FY83/84 accessions; and (c) administration of the revised predictor tests to a large sample of FY86/87 accessions and evaluation of their subsequent first-tour performance. The products from this comprehensive research undertaking have application both in present Army personnel operations and in continuing efforts to improve the selection and classification system.

EDGAR M. JOHNSON
Technical Director

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED
PERSONNEL: FINAL REPORT ON PROJECT A

EXECUTIVE SUMMARY

Requirement:

Project A was a comprehensive U.S. Army program to develop an improved
system to select and classify enlisted personnel. The system encompasses
675,000 persons and several hundred Military Occupational Specialties (MOS).
The objectives were to (a) validate existing selection measures against both
existing and project-developed criteria and develop new measures, (b) validate
early criteria (e.g., performance in training) as predictors of later criteria
(e.g., job performance) to improve assignment and promotion decisions, and
(c) determine the relative utility to the Army of different performance levels
across MOS.

Procedure:

With the Deputy Chief of Staff for Personnel as sponsor, work on the
long-term project was begun in 1982. In the first stage, relationships
between the scores applicants made on the Armed Services Vocational Aptitude
Battery (ASVAB) and their later performance in training and first-tour skill
tests were explored using file data for FY81/82 Army accessions.

The second stage was executed with FY83/84 accessions in 19 MOS,
selected as representative of the Army's 250+ entry-level MOS and accounting
for 45 percent of Army accessions. A preliminary battery of predictor mea-
sures (perceptual, spatial, temperament, interest, and biodata) was tested
with several thousand soldiers as they entered four MOS; revised versions were
field tested with nine MOS. The resulting predictor battery and a comprehen-
sive set of school knowledge tests, job knowledge tests, hands-on tests, and
performance ratings were administered in 1985 to 9,500 soldiers in 19 MOS in
the "Concurrent Validation." The results were used to analyze the components
of first-tour performance on the job (General Soldiering Skills, MOS-Specific
Skills, Leadership/Effort, Personal Discipline, Military Bearing/Physical
Fitness), and to compare the validities of the current ASVAB composites and
the added predictor measures for predicting job performance.

In the third stage, known as the "Longitudinal Validation," the revised
predictor measures were used to test more than 49,000 recruits at the time
they entered 21 MOS in FY86/87. Soldiers from this sample were tested on
their performance during training and are being tested during their first tour
on the job. Soldiers from the FY83/84 sample were also tested on their
second-tour performance.

Findings:

Project A products are of two general kinds: products for the "science" (personnel research) and products for the organization (the Army). However, many products are useful for both fields.

(1) Comprehensive reviews exist, in technical report form, of all validity evidence pertaining to selection and classification for skilled jobs. These are the most comprehensive reviews of this type ever done.

(2) Using much more comprehensive samples than ever before, new ASVAB aptitude area composites have been developed that are firmly data based and empirically defensible. The analyses involving ASVAB have resulted in a much clearer idea of its factor structure, of what the factors are measuring, and of its strengths and limitations.

(3) The question of whether ASVAB does or does not predict job performance (in addition to training performance) has been answered definitively in the affirmative. The Army and the Department of Defense are now in a more informed position to support their quality goals.

(4) A set of new experimental tests has been developed to measure noncognitive, psychomotor, perceptual, and cognitive characteristics that are not now measured by the ASVAB. The scope of Project A made it possible to examine virtually the entire domain of selection information, sample from it, and investigate the basic incremental validity produced by each major piece of information.

(5) Within the limits of the Concurrent Validation design, the incremental validity of appropriate ABLE temperament scales for predicting the "will do" components of performance has been demonstrated. The potential of the AVOICE interest scales for differentially predicting "can do" performance in combat vs. technical vs. administrative support MOS has been established.

(6) Much has been learned about the nature of performance in entry-level skilled jobs (e.g., first-tour MOS). We now have a much clearer idea of what major factors constitute performance and how they can be measured.

(7) The Project A job/task analysis procedures worked well and can be used by the Army in the future to develop training curricula, Skill Qualification Test content, performance measures, and field exercises.

x

(8)   Advanced Individual Training (AIT) achievement measures have been developed for 21 MOS. The training measures will allow a determination of whether training performance predicts job performance, and whether it does so differentially for different groups of trainees (race, gender), and different groups of MOS (combat, combat support, combat service support).

(9)   The package of rating scale administration procedures can be used in future personnel research in the Army. A major effort in Project A was to develop an effective and efficient set of procedures for administering performance rating scales to large numbers of people.

(10)   The data indicate that supervisor ratings of subordinate performance have considerable construct validity if a careful measurement procedure is followed. Supervisors seem to assess both the technical performance of individuals and their general dependability/motivation at the same time.

(11)   One very real, and very important, product is the Project A data base itself. It is by orders of magnitude the largest and most completely documented personnel research data base in existence.

Utilization of Findings:

The Project A tests for predicting and measuring training and job performance are being used in both current and long-range research programs that are expected to make the Army more effective in matching the requirements for first- and second-tour enlisted manpower with the personnel resources that are available to the Army. Additionally, Project A findings have already been used to make substantial improvements to the existing selection and classification system.

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED
PERSONNEL: FINAL REPORT ON PROJECT A

CONTENTS

CONTENTS (Continued)

CONTENTS (Continued)

LIST OF TABLES

Page

Page

LIST OF FIGURES

CONTENTS (Continued)

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION
OF ARMY ENLISTED PERSONNEL: FINAL REPORT ON PROJECT A
Chapter 1
INTRODUCTION

AN OVERVIEW OF PROJECT A

The Army annually contacts 400,000 to 500,000 young men and women, selects 90,000 to 130,000 of them, and assigns each individual to one of some 275 occupational specialties. Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel, and Project B: An Enlisted Personnel Allocation System, were designed to provide the greatest possible increase in overall performance and readiness that can be obtained from improved selection, classification, and allocation of enlisted personnel. These two research programs provided an integrated examination of performance measurement, selection/classification, supply and demand parameters, and allocation procedures to enable the Army to attempt optimizing the achievement of multiple personnel management goals (e.g., increase performance and decrease attrition).

The broad responsibilities of Project A were to develop:

● A comprehensive set of new predictor measures, following on valida-
tion of existing measures.

● Multiple measures of job performance, against which selection/
classification measures can be evaluated.

● Accurate estimates of the predictability of future performance.

● Decision rules for selection/classification at enlistment and
reenlistment to optimize individual and system performance.

● A way of evaluating the relative utility to the Army of different
performance levels across MOS.

Origins of Project A

The impetus for Project A came from the practical, professional, and legal need to demonstrate the validity of the Armed Services Vocational Aptitude Battery (ASVAB) and other selection variables for predicting job performance. Much of the existing validity data was based on using training measures as criteria.

In response to Army, Congressional, and professional requirements, the Army Research Institute (ARI) began in 1980 to develop a major new research program for personnel selection, classification, and allocation. The basic requirement was to demonstrate the validity of the ASVAB as a predictor of both training and on-the-job performance. In reviewing the design needed to meet that requirement, the concept of a larger project began to emerge. With only a moderate amount of additional resources, new selection/ classification measures in the perceptual, psychomotor, interest, temperament, and biodata domains could be evaluated as well. In addition, a longitudinal research data base could be developed, linking soldiers' performance on a variety of variables from enlistment, through training, first-tour assignments,

reenlistment decisions, and for some, to their second tour. Finally, the validation data could be the basis for new methods of allocating personnel, and making near-real-time decisions on the best match between characteristics of an individual enlistee or reenlistee and requirements of available Army Military Occupational Specialties (MOS).

To address the selection and classification portion of the effort, solicitation MDA 903-81-12-R-0158 "Project A: Development and Validation of Army Selection and Classification Measures" was issued 21 October 1981. This document can be viewed as the official starting point of Project A. The research program was intended to bring together Army and contractor research personnel in a combined effort to meet the Army's requirements for improving the processes and programs for selecting and classifying enlisted personnel. In the solicitation, the Army psychologists mapped out a comprehensive 7-year research program to provide the instrumentation and data necessary to implement a state-of-the-art selection and classification system for all enlisted personnel. (To provide background, a description of the present Army personnel system is included as Appendix A.)

While the contract solicitation process was ongoing, the new Manpower and Personnel Research Laboratory was created within ARI, and Dr. Joyce L. Shields was chosen as director. To accommodate the substantial in-house portion of Project A, the Selection and Classification Technical Area was established, with Dr. Newell K. Eaton as chief.

Formation of the Consortium

In anticipation of the solicitation, the Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), and Personnel Decisions Research Institute (PDRI) formed a consortium to develop a research proposal to meet the requirements of the forthcoming "Development and Validation of Army Selection and Classification Measures" Request for Proposal (RFP). It was agreed that HumRRO, as prime contractor, would assume responsibilities for overall contract management, technical direction, planning, and reporting. The proposal was submitted in January 1982 and the contract was awarded to the HumRRO-AIR-PDRI consortium 30 September 1982.

Project Outline

The overall purpose of Project A was to enhance the Army's ability to accomplish its peacetime and mobilization missions through improved matching of individuals to Military Occupational Specialties. Specifically, Project A was to

(1) Validate existing selection measures against both existing and project-developed criteria, the latter to include both Army-wide performance measures based on newly developed rating scales and direct measures of MOS-specific task performance.

(2) Develop and validate new and/or improved selection and classification measures.

(3) Validate proximal criteria, such as performance in training, as predictors of later criteria, such as job performance ratings, so that more informed reassignment and promotion decisions can be made throughout the individual's tour.

(4) Determine the relative utility to the Army of different performance levels across MOS.

(5) Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

The Statement of Work required that Project A be designed as one integrated project organized into five major tasks:

Task 1. Validation. Task 1 had two major components. The first was to develop and maintain the data base and provide the analytic procedures to determine the degree to which performance in Army jobs is predictable from some combination of new or existing measures. The second component was to conduct the appropriate analyses to determine whether the existing set of predictors, new predictors, or some combination of new and existing predictors has utility over and above the present system.

Task 2. Develop Predictors of Job Performance. A large proportion of the efforts of the Armed Services in this regard have been concentrated on improving the ASVAB, which is now a well-researched, valid measure of general cognitive abilities. However, many critical Army tasks appear to require psychomotor and perceptual skills for their successful performance. Further, neither biodata nor motivational variables were comprehensively evaluated. The objectives of Task 2 were to develop a broad array of new and improved selection measures and to administer them to three major validation samples. A critical aspect of this task was to be the demonstration of the incremental validity added by new predictors.

Task 3. Measurement of School/Training Success. The objective of Task 3 was to derive school and training performance indexes that could be used (a) as criteria against which to validate the initial predictors, and (b) as predictors of later job performance.

Task 4. Assessment of Army-wide Performance. In contrast to performance measures that may be developed for a specific Army MOS, Task 4 was to develop measures that could be used across all MOS (i.e., Army-wide). The intent was to develop measures of first- and second-tour job performance against which all Army enlisted personnel could be measured. A major objective was to develop a model of soldier effectiveness that specifies the major dimensions of an individual's contribution to the Army as an organization. Another important objective of Task 4 was to develop a procedure that could be used to scale the utility of levels of performance.

Task 5. Develop MOS-Specific Performance Measures. Task 5 was focused on developing reliable and valid measures of specific job task performance for a selected set of MOS. This task had three major components: job analysis, construction of job performance measures, and construct validation of the new measures. While only a subset of MOS were analyzed during this project, the

3

Army may in the future wish to develop job performance measures for a larger number of MOS. For this reason, the methodology was to apply to all Army MOS.

Initial Project Organization

The initial project organization is shown in Figure 1-1. The principal consortium task scientists are shown, with their respective organizations, in the lower row. The principal ARI scientists are shown in the upper row. Consortium and ARI scientists carried out research activities both independently and jointly. ARI scientists also had the administrative role of contract oversight.

We include this diagram only to show the matching of contractor and ARI staff and to illustrate the form of the project management and contract review structure. There were of course a number of personnel changes over the life of the project.

The Advisory Group Structure

A project of this scale would have to maintain close and active coordination with the other military departments and the Department of Defense, as well as remain consistent with other ongoing research programs being conducted by the other Armed Services. The project also needed a mechanism for assuring that the research program met the highest standards for scientific quality. Finally, a method was needed to receive feedback from senior officers on priorities and objectives, as well as to identify current problems. An effective mechanism for meeting these needs was deemed to be a structure of advisory groups.

Figure 1-2 shows the structure and membership of the Governance Advisory Group (GAG), which is made up of the Scientific Advisory Group (SAG), Interservice Advisory Group (ISAG), and Army Advisory Group (AAG) components.

The SAG was comprised of nationally recognized authorities in psychometrics, experimental design, sampling theory, utility analysis, applied research in selection and classification, and the conduct of psychological research in the Army environment. It is perhaps indicative of the substance and success of Project A that all members of the Scientific Advisory Group remained with the project from its beginning to the end.

The ISAG was comprised of the Laboratory Directors for applied psychological research in the Army, Air Force, and Navy, and the Director of Accession Policy from the Office of Assistant Secretary of Defense for Manpower and Reserve Affairs. The AAG included representatives from the Office of Deputy Chief of Staff for Personnel (DCSPER), Office of Deputy Chief of Staff for Operations (DCSOPS), Training and Doctrine Command (TRADOC), Forces Command (FORSCOM), and U.S. Army Europe (USAREUR).

4
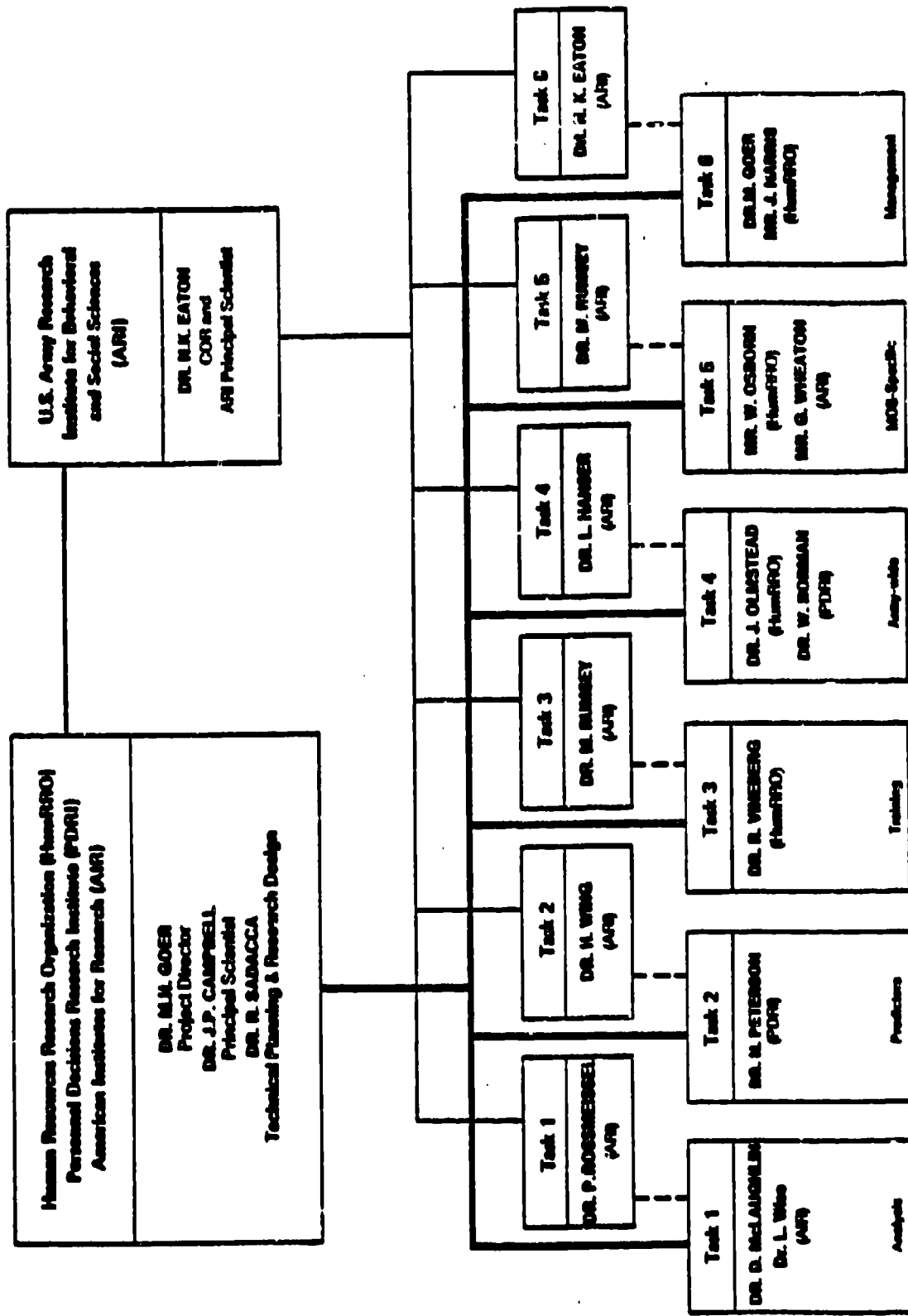
Figure 1-1. Initial Project A Organization.

5

Governance Advisory Group

MG N. SCHWARZKOPF
(CHAIRMAN)

SCIENTIFIC ADVISORS
- DR. P. BOBKO
- DR. T. COOK
- DR. M. HAKEL (Chair)
- DR. L. HUMPHREYS
- DR. R. LINN
- DR. M. TENOPYR
- DR. J. UHLANER

U.S. ARMY ADVISORS
- MG W.G. O'LEKSY
- MG R. BROWN
- DEPT. J.W. FOSS
- MG G.E. LUCK

INTERSERVICE ADVISORS
- DR. G.T. SICILIA (OSD)
- DR. J.L. SHIELDS (Army)
- DR. M.F. WISKOFF (Navy)
- COL. J.P. AMOR (Air Force)

Coordination and Support

U.S. Army Research Institute for Behavioral and Social Sciences (ARI)

DR. N.K. EATON

Human Resources Research Organization (HumRRO)

Personnel Decisions Research Institute (PDRI)

American Institutes For Research (AIR)

Figure 1-2. Initial Project A Governance Advisory Group.

6

Development of the Research Plan and the Integrated Master Plan

The first 6 months of the project were spent in planning, documenting, reviewing, modifying, and redrafting of research plans, troop support, administrative support, and budgetary plans, as well as in execution of initial research efforts. Drafts of the plans were provided to the SAG and ISAG. The culminating review was conducted in April 1983 by the Army Advisory Group, with representatives from the Scientific and Interservice Advisory Groups. The research program was endorsed by all three components of the GAG, and in May 1983, ARI issued Research Report 1332, Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Project A Research Plan.

An Outline of the Project A Research Plan

The Project A Research Plan spoke to the specific operational and scientific outcomes that would flow from the project.

Operational Objectives

The operational objectives were to --

(1)   Develop new measures of job performance that can be used as criteria against which to validate selection/classification measures.

(2)   Validate existing selection measures against both existing and project-developed criteria.

(3)   Develop and validate new selection and classification measures.

(4)   Develop a utility scale for different performance levels across MOS.

Research Objectives

The research objectives were to --

(1)   Identify the constructs that constitute the universe of information available for selection/classification into entry-level skilled jobs.

(2)   Develop a general model of performance for entry-level skilled jobs.

(3)   Investigate the construct validity of the "method" variance in job performance measures.

(4)   Estimate the value of different levels of job performance.

(5)   Estimate the degree of differential prediction across (a) major domains of predictor information (e.g., abilities, temperament, interests), (b) major factors of job performance, and (c) different types of jobs.

(6)   Determine the extent of differential prediction across racial and
        gender groups for a systematic sample of individual differences,
        performance factors, and jobs.

Research Design

The overall design of Project A used two predictive and one concurrent
validation on two major troop cohorts (1983/1984 accessions and 1986/1987
accessions), and one file data validation on the 1981/1982 cohort. That is,
in addition to collecting data from new samples, the project made use of
existing file data for 1981 and 1982 accessions. Data from the accessions and
Enlisted Master Files (EMF) were edited and merged into the Longitudinal
Research Data Base (LRDB). A schematic of the data collection plan is shown
in Figure 1-3.

The logic of the design was straightforward. Existing file data on the
81/82 cohort would provide an early opportunity to modify the existing
operational selection and classification decision rules; and in fact, the file
data analyses were used to recommend changes in the composition of the ASVAB
Aptitude Area composites. The 83/84 cohort provided the first opportunity to
obtain data using new predictor and performance measures. A "preliminary"
battery of predominantly off-the-shelf tests provided new predictor data on
soldiers in four MOS (05C, 19E/K, 63B, 71L). These data together with an
exhaustive literature search, job analysis information, and multiple expert
panel reviews provided the information to construct a more tailored trial
battery which was administered concurrently with a variety of training, Army-
wide, and MOS-specific performance measures in 1985 to the 1983/84 cohort.

The refinement of these measures resulted in the Experimental Predictor
Battery which was administered to a longitudinal sample from the FY86/87
cohort. The job performance criterion measures were administered to this
cohort during late 1988. In addition, at this same time second-tour per-
formance measures were developed for and administered to the FY83/84 cohort as
part of a longitudinal followup of that sample into its second tour.

MOS and Sample Selection

The overall objective in generating the samples was to maximize the
validity and reliability of the information to be gathered, while at the same
time minimizing the time and costs involved. While costs are a function of
the numbers of people in the sample, they are also influenced by the relative
difficulty involved in locating and assembling the people in a particular
sample.

The sampling plan itself incorporated two principal considerations.
First, a sample of MOS was selected from the universe of possible MOS. Then,
the required sample sizes of enlisted personnel within each MOS were speci-
fied. Because Project A was developing a system for a population of jobs
(MOS), the MOS are the primary sampling units.

There is a trade-off in the allocation of resources between the number
of MOS researched and the number of subjects tested within each MOS: the more
MOS are investigated, the fewer subjects per MOS can be tested, and vice
versa. Cost and statistical reliability considerations dictated that 19 MOS
could be studied. The new predictors (from Task 2) as well as the school and

8
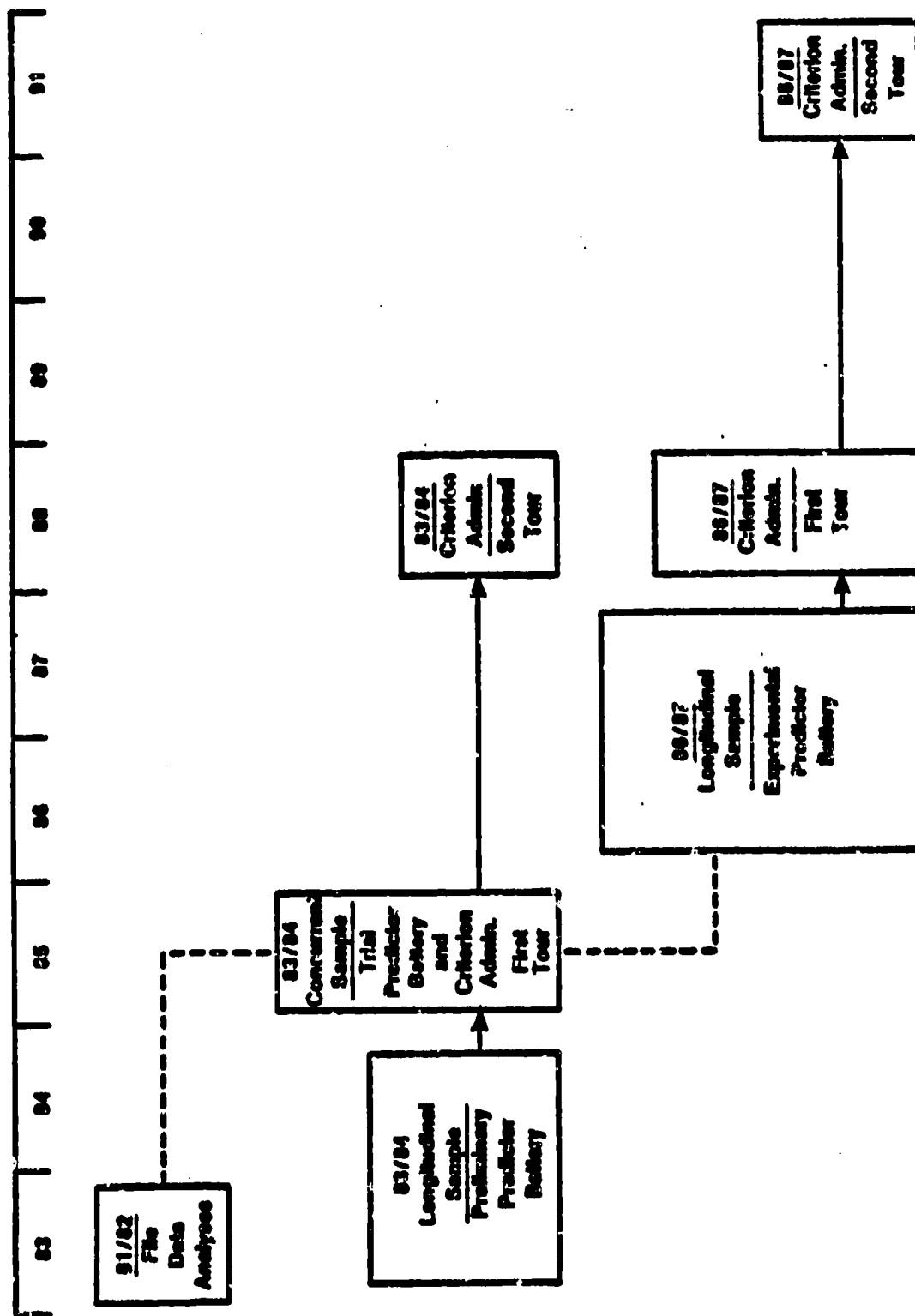
Figure 1-3. The overall research design for Project A.

9

Army-wide performance measures (of Tasks 3 and 4) were administered to all 19. For nine of the 19 MOS, the MOS-specific performance measures developed in Task 5 were also administered; the nine MOS were chosen to provide maximum coverage, given certain statistical constraints, of the total array of knowledge, ability, and skill requirements of Army jobs.

The selection of the sample of 19 MOS proceeded through a series of stages. An initial sample of MOS was drawn on the basis of the following considerations:

(1) High-density MOS that would provide sufficient sample sizes for statistically reliable estimates of new predictor validity and differential validity across racial and gender groups.

(2) Representative coverage of the aptitude areas measured by the ASVAB area composites.

(3) High-priority MOS (as rated by the Army in the event of a national emergency).

(4) Representation of the Army's designated Career Management Fields (CMF).

(5) Representation of the jobs most crucial to the Army's mission.

A further indirect indication of the mix of job skills represented in the sample is in the range of ASVAB composites and component subtests pertinent to each MOS. The ASVAB subtests are Word Knowledge (WK), Paragraph Comprehension (PC), Arithmetic Reasoning (AR), Numerical Operations (NO), General Science (GS), Mechanical Comprehension (MC), Math Knowledge (MK), Electronics Information (EI), Coding Speed (CS), and Auto-Shop Information (AS). The WK and PC subtest raw scores are summed to create an additional Verbal (VE) subtest. The composites, combinations of subtests to characterize aptitude areas, are Clerical (CL), Combat (CO), Electronics (EL), Field Artillery (FA), General Maintenance (GM), Mechanical Maintenance (MM), Operators/Food (OF), Surveillance and Communication (SC), and Skilled Technical (ST).

All subtests and all but one (Electronics) of the nine composites were represented in the 18 MOS initially selected. Consequently, a 19th MOS (27E) was chosen to represent the EL aptitude composite. The composition of the sample was also examined from the perspective of mission criticality by comparing it with a list of 42 MOS identified by the Army as high priority for mobilization training.[1] This initial set of 19 MOS represent 19 of the Army's 30 CMF. Of the 11 CMF not represented, two are classified (CMF 96 and 98), two (CMF 33 and 74) had fewer than 500 FY81 accessions, and seven (CMF 23, 28, 29, 79, 81, 84, and 74) had fewer than 300 FY81 accessions. The initial MOS set included only 5 percent of Army jobs but 44 percent of the soldiers recruited in FY81. Similarly, of the 15 percent women in the Army, 44 percent are represented in the sample.

---

[1]ODCSOPS (DAMO-ODM), DF, 2 Jul 82, Subject: IRR Training Priorities.

Guidance from the Scientific Advisory Group led to further refinement of the MOS sample. A cluster analysis of expert ratings of MOS similarity was made, and the initial sample was reviewed by the Governance Advisory Group.

To obtain data for empirically clustering MOS on the basis of their task content similarity, a brief job description was generated for each of 111 MOS from the job activities described in AR 611-201.[c] The sample of 111 MOS included the 84 largest MOS (300 or more new job incumbents yearly) plus an additional 27 selected randomly but proportionately by CMF. Each job description was limited to two sides of a 5x7 index card.

Members of the contractor research staff and ARI Army officers (N = 25), serving as expert judges, sorted the sample of 111 job descriptions into homogeneous categories based on perceived similarities and differences in the described job activities. The similarity data were clustered and used to check the representativeness of the initial sample of 19 MOS. (That is, did the 19 MOS include representatives from all the major clusters of MOS derived from the similarity scaling?) On the basis of these results and guidance received from the Governance Advisory Group, two MOS that had been selected initially were replaced.

The initial sample of 19 MOS resulting from the above procedures is shown in Table 1-1. The subsample of nine MOS to which the MOS-specific

Table 1-1

Initial List of Project A Military Occupational Specialties (MOS)

	BATCH A[a]		BATCH Z
05C	Radio Teletype Operator[b]	12B	Combat Engineer
11B	Infantryman	16S	MANPADS Crewman
13B	Cannon Crewman	27E	TOW/Dragon Repairer
19E	Tank Crewman	51B	Carpentry/Masonry Specialist
63B	Vehicle & Generator Mechanic Specialist	54E	Chemical Operations Specialist
64C	Motor Transport Operator	55B	Ammunition Specialist
71L	Administrative Specialist	67N	Utility Helicopter Repairer
91A	Medical Care Specialist	76W	Petroleum Supply Specialist
95B	Military Police	76Y	Unit Supply Specialist
		94B	Food Service Specialist

[a] MOS-specific criterion measures were administered in these MOS.
[b] MOS 05C later became MOS 31C.

[c] Army Regulation 611-201, Enlisted Career Management Fields and Military Occupational Specialties.

criterion measures were administered is shown as Batch A. During the course of the project, some MOS changed names or numbers, some were added or deleted because requirements changed. The MOS lists in the report reflect these changes as they occurred. One of the original MOS (76W) was deleted and three MOS (19K, 20E, and 96B) were added, making a total of 21 MOS in the sample during the later stages of the research.

ORGANIZATION OF THIS PROJECT A REPORT

Given the basic design just described, the remainder of this report summarizes the substantive work of Project A from October 1982 through March 1990. Since Project A was large in scope, the summary is not short. The intent was to provide enough detail to permit a judgment about the thoroughness and appropriateness of the work done at each step.

The content of the summary was assembled from the FY83, FY84, FY85, FY86, FY87, and FY88 project annual reports, which in turn were based on very detailed technical reports, working papers, and convention papers on specialized topics. The full Bibliography of reports, papers, and products for the duration of Project A is included as Appendix B. The names of the people who worked on Project A are presented in Appendix C.

The major topics covered in this final report are:

- Development of new selection/classification (predictor) tests.

- Development of new measures of training and job performance.

- Concurrent Validation procedure.

- Development of basic prediction and criterion scores.

- Results of the Concurrent Validation.

- Development of differential weights for the major components of job performance.

- The scaling of the utility of performance in entry-level jobs.

- Job analyses and criterion development for second-tour MOS.

- Samples and procedures for the Longitudinal Validation.

The final chapter of the report discusses the Project A research in the context of selection and classification history, and highlights its products and findings in terms of both basic and applied research concerns and goals.

Chapter 2
PREDICTOR DEVELOPMENT

SELECTION OF VARIABLES

The overall goal of predictor development in Project A was to construct an experimental test battery that would, when combined with ASVAB, yield the maximum increment in selection/classification validity for the entire system. That is, what new tests should be used in conjunction with ASVAB to increase the aggregate accuracy of selection and classification decisions over all MOS in the enlisted personnel system? Approximately 280 MOS now use ASVAB for such decisions.

Given this overall goal, the Project A research staff adopted a very comprehensive approach that tried to (a) define the population of potentially useful variables; (b) describe its latent structure; (c) sample constructs from this population that had the highest probability of meeting the goals of the project; (d) construct operational measures of these variables; (e) pilot test, field test, and revise the new measures; (f) analyze their empirical covariance structure; and (g) determine their predictive validities, and specify the optimal decision rules for using the new tests to maximize predicted performance and/or minimize attrition. The major steps that were taken to execute this approach are described in this chapter (also see Peterson, 1986; Peterson et al., 1987).

Review of Selection/Classification Literature

The overriding purpose of the literature review was to gain maximum benefit from earlier research that was even remotely relevant for the jobs in the Project A job population. The search was conducted in late 1982 and early 1983 (i.e., FY83) by three teams of project staff.

Several computerized searches of all relevant data bases resulted in identification of more than 10,000 sources. In addition, reference lists were solicited from recognized experts, annotated bibliographies were obtained from military research laboratories, and the last several years' editions of relevant research journals were examined, as were more general sources such as textbooks, handbooks, and appropriate chapters in the Annual Review of Psychology.

The references identified as relevant were obtained, reviewed, and summarized using a standardized report protocol of seven sections: description of predictor, reliability, norms/descriptive statistics, correlations with other predictors, correlations with criteria, adverse impact/differential validity/test fairness, and reviewer's recommendations (about the usefulness of the predictor). Each predictor was tentatively classified into an initial, working taxonomy of predictor constructs (based primarily on the taxonomy described in Peterson and Bownas, 1982).

13

Literature Search Results

The literature search was used in two major ways. First, three working documents were written, one for each of three areas: cognitive/perceptual abilities, psychomotor/perceptual abilities, and non-cognitive predictors (including temperament or personality, vocational interest, and biographical data variables). These documents summarized the literature with regard to critical issues, suggested the most appropriate organization or taxonomy of the constructs in each area, and summarized the validities of the various measures for different types of job performance criteria. (These documents were subsequently issued as Hough, 1986; McHenry & Rose, 1986; Tcquam, Corpe, & Dunnette, 1990.)

Second, the predictors identified in the review were subjected to further scrutiny to (a) select tests and inventories to make up the Preliminary Battery, and (b) select the "best bet" predictor constructs to be used in the "expert judgment" research activity.

Screening of Predictors

An initial list was compiled of all predictor measures that seemed even remotely appropriate for Army selection and classification. This list was then screened by eliminating measures according to several "knockout" factors: (a) measures developed for a single research project; (b) measures designed for a narrowly specified population/occupational group (e.g., pharmacy students); (c) measures targeted toward younger age groups; (d) measures requiring unusually long testing times; (e) measures requiring difficult or subjective scoring; and (f) measures requiring individual administration.

Application of the knockout factors resulted in a second list of candidate measures that served as the final selection of constructs to be included in the "expert judgment." This research was designed to use expert judgment to estimate the potential validity of each relevant construct, if it were reliably measured. Schmidt, Hunter, Croll, and McKenzie (1983) have shown that pooled expert judgments, obtained from experienced personnel psychologists, have considerable accuracy for estimating the validity of tests in actual, empirical, criterion-related validity research.

Expert Forecasts of Predictor Construct Validities

Peterson and Bownas (1982) provide a complete description of the methodology which has been used successfully by Bownas and Heckman (1976), Peterson, Houston, Bosshardt, and Dunnette (1977), Peterson and Houston (1980), and Peterson, Houston, and Rosse (1984) to identify predictors for the jobs of firefighter, correctional officer, and entry-level occupations (clerical and technical), respectively. Descriptive information about a set of predictors and the job performance criterion variables is given to "experts" in personnel selection and classification. These experts estimate the relationships between predictor and criterion variables by rating or directly estimating the value of the correlation coefficients.

The result is a matrix with predictor and criterion variables as the columns and rows, respectively. Cell entries are experts' estimates of the degree of relationship between the particular predictors and various criteria. The interrater reliability of the experts' estimates is checked first. If the

14

estimate is sufficiently reliable (previous research shows values in the .80 to .90 range for about 10 to 12 experts), the matrix of predictor-criterion relationships can be analyzed and used in a variety of ways. For example, by correlating the rows of the matrix the covariances between criteria can be estimated, and by correlating the columns the covariances between predictors can be estimated on the basis of the profiles of their estimated relationships with the criteria. The covariances can then be factor analyzed to identify clusters of predictors within which the measures are expected to exhibit similar patterns of correlations with different performance components. Similarly, the criterion covariances can be examined to identify clusters of criteria predicted by a common set of predictors.

Such procedures helped in identifying redundancies and overlap in the predictor set. The clusters of predictors and of criteria are an important product for a number of reasons. First, they provide an efficient and organized means of summarizing the data generated by the experts. Second, the summary form permits easier comparison with the results of meta-analyses of empirical estimates of criterion-related validity coefficients. Third, these clusters provide a model or theory of the predictor-criterion performance space.

Method

For Project A, the experts were 35 industrial, measurement, or differential psychologists with experience and knowledge in personnel selection research and/or applications.

The previous reviews of the population of constructs had identified a basic list of 53 variables, and materials describing each of these variables were prepared. The procedure used to identify criterion variables was based on the job descriptions of the sample of 111 MOS that had been previously clustered by job experts as part of the MOS sample selection. Criterion categories were developed by reviewing the descriptions to determine common job performance activities.

After common elements in the 23 clusters were identified, additional categories were identified to cover unique aspects of jobs in the sample of 111. Most of the 53 performance component categories applied to several jobs, and most of the jobs were characterized by activities from several categories. The second type of criterion variable was a set that described performance in initial Army training as defined in archival records and interviews with trainers. The final set of criterion variables consisted of the general performance categories defined by the behavioral dimensions developed as part of Task 4. In all, 72 possible criterion constructs were defined for use in the expert judgment task.

Each judge estimated the true validity of each predictor for each criterion (i.e., criterion-related validity corrected for such artifacts as range restriction and reliability, and unaffected by variation in sample sizes). All judges completed the task during the first week of October 1983.

When averaged across raters, the reliability of the mean estimated cell validities was .96. Factor analyses were based on these cell means. The most pertinent for purposes of this summary report concerns the analysis of the predictor profiles.

Eight interpretable factors were named: I, Cognitive Abilities; II, Visualization/Spatial; III, Information Processing; IV, Mechanical; V, Psychomotor; VI, Social Skills; VII, Vigor; VIII, Motivation/Stability. These eight factors appeared to be composed of 21 clusters, and the hierarchical structure is shown in Figure 2-1.

Variables for measurement were sampled from the hierarchy on the basis of (a) a careful review of the empirical literature within each category, (b) visits to all major military personnel research stations, (c) on-site observations of individuals during field exercises in the combat specialties, and (d) a multistage review of all available information by the project staff and the Scientific Advisory Group.

Identification of Pilot Trial Battery Measures

In March 1984, a formal In Progress Review (IPR) meeting was held to decide on the measures to be developed for the Pilot Trial Battery. Information from the literature review, expert judgments, initial analyses of the Preliminary Battery, and the first three phases of computer battery development was presented and discussed. The Project A staff made recommendations for inclusions of measures and these were evaluated and revised. Figure 2-2 shows the results of that deliberation process.

This set of recommendations constitutes the initial array of predictor variables for which measures would be constructed and then submitted to a series of pilot tests and field tests, with revisions being made after each phase.

PREDICTOR DEVELOPMENT: COGNITIVE PAPER-AND-PENCIL MEASURES

Development of measurement operations for the high-priority constructs considered the following issues: (a) a definition of the target cognitive ability; (b) the target population or target MOS for which the measure is hypothesized to most effectively predict success; (c) published tests that served as markers for each new measure; (d) intended level of item difficulty; and (e) type of test (i.e., speed, power, or a combination).

Brief descriptions of the individual tests, as initially designed, are given below, along with an explanation of the constructs the tests are intended to represent.

CONSTRUCTS	CLUSTERS	FACTORS
1. Verbal Comprehension 5. Reading Comprehension 19. Ideational Fluency 18. Analogical Reasoning 21. Omnibus Intelligence/Aptitude 22. Word Fluency	A. Verbal Ability/ General Intelligence	
4. Word Problems 9. Inductive Reasoning: Concept Formation 10. Deductive Logic	B. Reasoning	
2. Numerical Computation 3. Use of Formula/Number Problems	C. Number Ability	COGNITIVE ABILITIES
13. Perceptual Speed and Accuracy	H. Perceptual Speed and Accuracy	
49. Investigative Interests	U. Investigative Interests	
14. Rote Memory 17. Follow Directions	J. Memory	
18. Figural Reasoning 23. Verbal and Figural Closure	F. Closure	
6. Two-dimensional Mental Rotation 7. Three-dimensional Mental Rotation 8. Spatial Visualization 11. Field Dependence (Negative) 15. Place Memory (Visual Memory) 20. Spatial Scanning	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
24. Processing Efficiency 25. Selective Attention 26. Time Sharing	G. Mental Information Processing	INFORMATION PROCESSING
12. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests 51. Artistic Interests (Negative)	M. Realistic vs. Artistic Interests	
28. Control Precision 29. Rate Control 33. Arm-hand Steadiness 34. Aiming	I. Steadiness/Precision	
27. Multilimb Coordination 35. Speed of Arm Movement	O. Coordination	PSYCHOMOTOR
30. Manual Dexterity 31. Finger Dexterity 32. Wrist-finger Speed	K. Dexterity	
39. Sociability 52. Social Interests	Q. Sociability	SOCIAL SKILLS
50. Enterprising Interests	R. Enterprising Interests	
36. Involvement in Athletics and Physical Conditioning 37. Energy Level	T. Athletic Abilities/Energy	VIGOR
41. Dominance 42. Self-esteem	S. Dominance/Self-esteem	
43. Traditional Values 45. Conscientiousness 46. Non-delinquency 53. Conventional Interests	N. Traditional Values/Conventionality/ Non-delinquency	
44. Locus of Control 47. Work Orientation	Q. Work Orientation/Locus of Control	MOTIVATION/ STABILITY
38. Cooperativeness 40. Emotional Stability	P. Cooperation/Emotional Stability	

Figure 2-1. Hierarchical map of predictor space.

17

Final Priority*	Predictor Category	Pilot Trial Battery Test Names
Cognitive:		
7	Memory	(Short) Memory Test - Computer
6	Number	Number Memory Test - Computer
8	Perceptual Speed & Accuracy	Perceptual Speed & Accuracy - Computer
		Target Identification Test - Computer
4	Induction	Reasoning Test 1
		Reasoning Test 2
5	Reaction Time	Simple Reaction Time - Computer
		Choice Reaction Time - Computer
3	Spatial Orientation	Orientation Test 1
		Orientation Test 2
		Orientation Test 3
2	Spatial Visualization/Field Independence	Shapes Test
1	Spatial Visualization	Object Rotations Test
		Assembling Objects Test
		Path Test
		Maze Test
Non-Cognitive, Biodata/Temperament:		
1	Adjustment	
2	Dependability	
3	Achievement	
4	Physical Condition	ABLE (Assessment of Background
5	Potency	Life Experiences)
6	Locus of Control	
7	Agreeableness/Likeability	
1	Validity Scales	
Non-Cognitive, Interests:		
1	Realistic	
2	Investigative	
3	Conventional	AVOICE (Army Vocational Interest
4	Social	Career Examination)
5	Artistic	
6	Enterprising	
Psychomotor:		
1	Multilimb Coordination	Target Tracking Test 2 - Computer
		Target Shoot - Computer
2	Precision	Target Tracking Test 1 - Computer
3	Manual Dexterity	(None)

*Final priority arrived at via consensus of March 1984 IPR attendants.

Figure 2-2. Predictor categories discussed at IPR in March 1984, linked to subsequent Pilot Trial Battery test names.

Spatial Visualization - Rotation

Spatial visualization involves the ability to mentally manipulate components of two- or three-dimensional figures into other arrangements. The process involves restructuring the components of an object and accurately discerning their appropriate appearance in new configurations. This construct includes several subcomponents, two of which are rotation and scanning. The two tests developed to measure visual rotation ability are Assembling Objects and Object Rotation, involving three-dimensional and two-dimensional objects, respectively.

Assembling Objects Test. This test was designed to assess the ability to visualize how an object will look when its parts are put together correctly. This measure was intended to combine power and speed components, with speed receiving greater emphasis. Each item presents subjects with components or parts of an object. The task is to select, from among four alternatives, the one object that depicts the components or parts put together correctly. Published tests identified as markers for Assembling Objects include the Employee Aptitude Survey Space Visualization (EAS-5) and the Flanagan Industrial Test (FIT) Assembly.

Object Rotation Test. The initial version contained 60 items with a 7-minute time limit. The subject's task is to examine a test object and determine whether the figure represented in each item is the same as the test object, only rotated, or is not the same as the test object (e.g., flipped over). Published tests serving as markers for the Object Rotation measure include Educational Testing Service (ETS) Card Rotations, Thurstone's Flags Test, and Shephard-Metzler Mental Rotations.

Spatial Visualization - Scanning

The second component of spatial visualization ability is spatial scanning, which requires the subject to visually survey a complex field and find a pathway through it, utilizing a particular configuration. The Path Test and the Maze Test were developed to measure this component.

Path Test. The Path Test requires subjects to determine the best path or route between two points. Subjects are presented with a map of airline routes or flight paths. The subject's task is to find the "best" path or the path between two points that requires the fewest stops. Published tests serving as markers for construction of the Path Test include ETS Map Planning and ETS Choosing a Path.

Maze Test. The first pilot test version of the Maze Test contained 24 rectangular mazes, with four entrance points and three exit points. The task is to determine which of the four entrances leads to a pathway through the maze and to one of the exit points. A 9-minute limit was established.

Field Independence

This construct involves the ability to find a simple form when it is hidden in a complex pattern. Given a visual percept or configuration, field independence refers to the ability to hold the percept or configuration in mind so as to distinguish it from other well-defined perceptual material.

19

<u>Shapes Test</u>. The marker test is ETS Hidden Figures. The strategy for constructing the Shapes Test was to use a task similar to that in the Hidden Figures Test while ensuring that the difficulty level of test items was geared more toward the Project A target population. The test was to be speeded, but not nearly so much so as the Hidden Figures. At the top of each test page are five simple shapes; below these shapes are six complex figures. Subjects are instructed to examine the simple shapes and then to find the one simple shape located in each complex figure.

Spatial Orientation

This construct involves the ability to maintain one's bearings with respect to points on a compass and to maintain location relative to landmarks. It was not included in the list of predictor constructs evaluated by the expert panel, but it had proved useful during World War II, when the Army Air Forces (AAF) Aviation Psychology Program explored a variety of measures for selecting air crew personnel. Also, during the second year of Project A, a number of job observations suggested that some MOS involve critical job requirements of maintaining directional orientation and establishing location, using features or landmarks in the environment. Consequently, three different measures of this construct were formulated.

<u>Orientation Test 1</u>. Direction Orientation Form B (CP515B) developed by researchers in the AAF Aviation Psychology Program served as the marker for Orientation Test 1. Each test item presented subjects with six circles. In the test's original form, the first, or Given, circle indicated the compass direction for North. For most items, North was rotated out of its conventional position. Compass directions also appeared on the remaining five circles. The subject's task was to determine, for each circle, whether or not the direction indicated was correctly positioned by comparing it to the direction of North in the Given circle.

<u>Orientation Test 2</u>. Each item contains a picture within a circular or rectangular frame. The bottom of the frame has a circle with a dot inside it. The picture or scene is not in an upright position. The task is to mentally rotate the frame so that the bottom of the frame is positioned at the bottom of the picture. After doing so, one must then determine where the dot will appear in the circle. The original form of the test contained 24 items, and a 10-minute time limit was established.

<u>Orientation Test 3</u>. This test was modeled after another spatial orientation test, Compass Directions, developed in the AAF Aviation Psychology Program. Orientation Test 3 presented subjects with a map that includes various landmarks such as a barracks, a campsite, a forest, a lake. Within each item, subjects are provided with compass directions by indicating the direction from one landmark to another, such as "the forest is North of the campsite." Subjects are also informed of their present location relative to another landmark. Given this information, the subject must determine which direction to go to reach yet another structure or landmark. For each item, new or different compass directions are given.

Induction/Figural Reasoning

This construct involves the ability to generate hypotheses about principles governing relationships among several objects. Example measures of induction include the Employee Aptitude Survey Numerical Reasoning (EAS-6), ETS Figure Classification, Differential Aptitude Test (DAT) Abstract Reasoning, Science Research Associates (SRA) Word Grouping, and Raven's Progressive Matrices. These paper-and-pencil measures present subjects with a series of objects such as figures, numbers, or words. To complete the task, subjects must first determine the rule governing the relationship among the objects and then apply the rule to identify the next object in the series. Two different measures of the construct were developed for Project A.

Reasoning Test 1. The plan was to construct a test that was similar to the task appearing in EAS-6, Numerical Reasoning, but with one major difference: Items would be composed of figures rather than numbers. Reasoning Test 1 items present subjects with a series of four figures; the task is to identify from among five possible answers the one figure that should appear next in the series.

Reasoning Test 2. The ETS Figure Classification test, which served as the marker, requires subjects to identify similarities and differences among groups of figures and then to classify test figures into those groups. Items in Reasoning Test 2 were designed to involve only the first task. The test items present five figures. Subjects are asked to determine which four figures are similar in some way, thereby identifying the one figure that differs from the others.

PREDICTOR DEVELOPMENT: COMPUTER-ADMINISTERED TESTS

There were four phases of activities: (a) information gathering about past and current research in perceptual/psychomotor measurement and computerized methods of testing such abilities; (b) construction of a demonstration computer battery; (c) selection of commercially available microprocessors and peripheral devices, writing of software for testing several abilities using this hardware, and tryout of this hardware and software; (d) continued development of software, and the design and construction of a custom-made response pedestal.

Compared to the paper-and-pencil measurement of cognitive abilities, computerized measurement of psychomotor and perceptual abilities was in a relatively primitive state. Much work had been done in World War II using electromechanical apparatus, but relatively little work had occurred since then. Microprocessor technology held out the promise of improving measurement in this area, but the work was (and still is) in its early stages.

Development of Response Pedestal

Development of the computer-administered measures was in turn dependent upon development of the appropriate hardware and software. The portable microprocessor selected for use was modeled after the COMPAQ but the preliminary trials suggested that the use of a keyboard may provide an unfair advantage to subjects who have typing or data entry experience, so a separate response pedestal was designed and built.

21

This response pedestal is depicted in Figure 2-3. Note that it contains two joysticks (one for left-handed and one for right-handed subjects), two sliding resistors, a dial for entering demographic data such as age and social security number, two red buttons, three response buttons--blue, yellow, and white--and four green "home" buttons.

To begin a trial, the subjects must place their hands on the four green buttons. After the stimulus appears on the screen and the subject has determined the correct response, he or she must remove the preferred hand from the "home" buttons and press the correct response button. The "home" buttons serve two purposes. First, control is added over the location of the hands while the stimulus item is presented. Second, procedures involving these buttons are designed to assess two theoretically important components of reaction time measures--decision time and movement time.
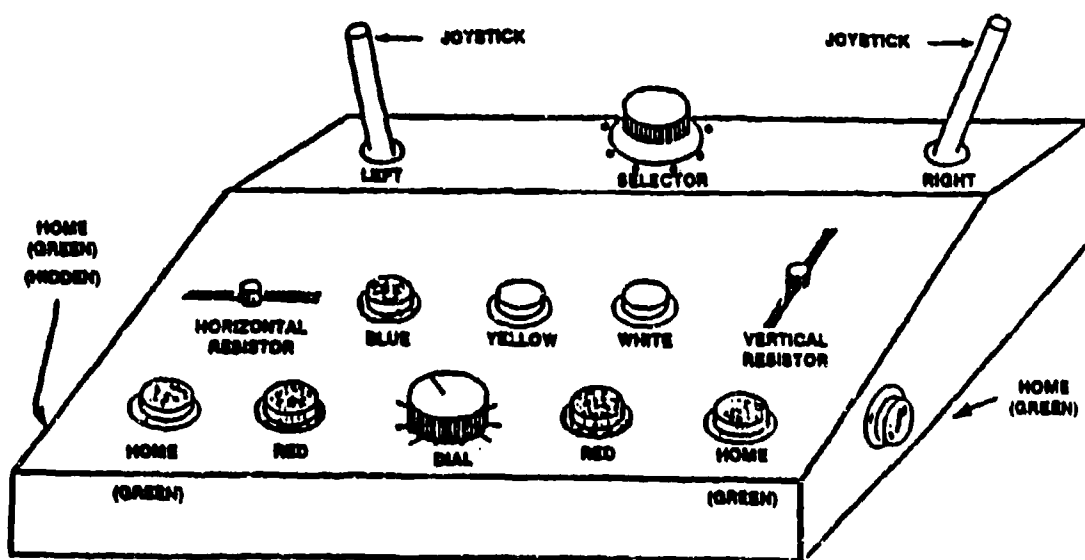
Figure 2-3. Response pedestal for computerized tests.

Test Development

Reaction Time (Processing Efficiency)

This construct involves speed of reaction to stimuli--that is, the speed with which a person perceives the stimulus independent of any time taken by the motor response component of the classic reaction time measures. It is intended to be an indicator of processing efficiency and includes both simple and choice reaction time.

Simple Reaction Time: RT Test 1. The basic paradigm for this task stems from Jensen's research involving the relationship between reaction time and mental ability (Jensen, 1982). On the computer screen, a small box appears. After a delay period (ranging from 1.5 to 3.0 seconds) the word YELLOW appears in the box. The subject must remove the preferred hand from the "home" buttons to strike the yellow key. The subject must then return both hands to the ready position to receive the next item.

Choice Reaction Time: RT Test 2. Reaction time for two response alternatives is obtained by presenting the word BLUE or WHITE on the screen. The subjects are instructed that, when one of these appears, they are to move the preferred hand from the "home" keys to strike the key that corresponds with the word appearing on the screen (BLUE or WHITE).

Short-Term Memory

This construct is defined as the rate at which one observes, searches, and recalls information contained in short-term memory.

Memory Search Test. The marker was a short-term memory search task introduced by S. Sternberg (1966, 1969) and the measure developed for Project A is similar. The first stimulus set appears and contains one, two, three, four, or five objects (letters). Following a display period of 0.5 or 1.0 second, the stimulus set disappears and, after a delay, the probe item appears. Presentation of the probe item is delayed by either 2.5 or 3.0 seconds and the subject must then decide whether or not it appeared in the stimulus set. If the item was present in the stimulus set, the subject strikes the white key. If the probe item was not present, the subject strikes the blue key.

Parameters of interest include the number of letters in the stimulus set, length of observation period, probe delay period, and probe status (i.e., the probe is either in the stimulus or not in the stimulus set). Subjects receive scores on the following measures:

> The Slope and Intercept obtained by regressing mean total reaction time (correct responses only) against item length. In terms of processing efficiency, the slope represents the average increase in reaction time with an increase of one object in the stimulus set. The intercept represents all other processes not involved in memory search, such as encoding the probe, determining whether or not a match has been found, and executing the response.

Percent Correct scores, used to identify subjects performing at very low levels which would preclude computation of the above scores.

The Grand Mean obtained by calculating the mean of the mean reaction time (correct responses only) for each level of stimulus set length (i.e., one to five).

Perceptual Speed and Accuracy

Perceptual speed and accuracy involves the ability to perceive visual information quickly and accurately and to perform simple processing tasks with the stimulus (e.g., make comparisons). This requires the ability to make rapid scanning movements without being distracted by irrelevant visual stimuli, and measures memory, working speed, and sometimes eye-hand coordination.

Perceptual Speed and Accuracy Test. Measures used as markers for the development of the computerized Perceptual Speed and Accuracy (PS&A) Test included the Employee Aptitude Survey Visual Speed and Accuracy (EAS-4) and the ASVAB Coding Speed. The computer-administered Perceptual Speed and Accuracy Test requires the ability to make a rapid comparison of two visual stimuli presented simultaneously and determine whether they are the same or different. Five different types of stimuli are presented: alpha, numeric, symbolic, mixed, and word. Within the alpha, numeric, symbolic, and mixed stimuli, the character length of the stimulus is varied. Four levels of character stimulus length are present: two, five, seven, and nine.

Target Identification Test. In this test, each item shows a target object near the top of the screen and three color-labeled stimuli in a row near the bottom of the screen. Examples are shown in Figure 2-4. The subject is to identify which of the three stimuli represents the same object as the target and to press, as quickly as possible, the button (blue, yellow, or white) that corresponds to that object. The objects shown are based on military vehicles and aircraft as shown on the standard set of flashcards used to train soldiers to recognize equipment presently being used by various nations. Several parameters were varied in the stimulus presentation. In addition to type of object, the position of the correct response (left or right side of the screen), the orientation of the target object (facing in the same direction as the stimuli or in the opposite direction), variation in the angle of rotation (from horizontal) of the target object, and the size of the target object were incorporated into the test.

Psychomotor Precision

This construct reflects the ability to make the muscular movements necessary to adjust or position a machine control mechanism. The ability applies both to anticipatory movements where the stimulus condition is continuously changing in an unpredictable manner and to controlled movements where stimulus conditions change in a predictable fashion. Psychomotor precision thus encompasses two of the ability constructs identified by Fleishman and his associates, control precision and rate control (Fleishman, 1967).

EXAMPLE 1.

TARGET
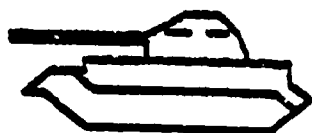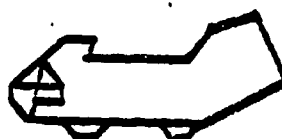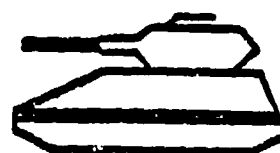
BLUE   YELLOW   WHITE

EXAMPLE 2.

TARGET

BLUE   YELLOW   WHITE

Figure 2-4.   Graphic displays of example items from the computer-
administered Target Identification Test.

25

Performance on tracking tasks is very likely related to psychomotor precision and, since tracking tasks are an important part of many Army MOS, development of psychomotor precision tests was made a high priority. The initial computer battery included two measures of this ability.

Target Tracking Test 1. This test was designed to measure control precision, and the AAF Rotary Pursuit Test served as a model. For each trial, subjects are shown a path consisting entirely of vertical and horizontal line segments. At the beginning of the path is a target box, and centered in the box are crosshairs. As the trial begins, the target starts to move along the path at a constant rate of speed. The subject's task is to keep the crosshairs centered within the target at all times. The subject uses a joystick, controlled with one hand, to control movement of the crosshairs.

Item parameters include the speed of the crosshairs, the maximum speed of the target, the difference between crosshairs and target speeds, the total length of the path, the number of line segments comprising the path, and the average amount of time the target spends traveling along each segment.

Two kinds of scores were investigated: (a) tracking accuracy and (b) improvement in tracking performance. Two accuracy measures were investigated, time on target and distance from the center of crosshairs to the center of the target. The test program computes the distance from the crosshairs to the center of the target several times each second, and then averages these distances to derive an overall accuracy score for that trial. Subsequently, to remove positive skew, each trial score was transformed by taking the square root of the average distance. These trial scores were then averaged to determine an overall tracking accuracy score.

Target Shoot Test. This test was modeled after several compensatory and pursuit tracking tests used by the AAF in the Aviation Psychology Program (e.g., the Rate Control Test). For the Target Shoot Test, a target box and a crosshairs appear in different locations on the computer screen. The target moves about the screen in an unpredictable manner, frequently changing speed and direction. The subject controls movement of the crosshairs via a joystick and the task is to move the crosshairs into the center of the target, and to "fire" at the target. The score is the distance from the center of the crosshairs to the center of the target.

Several item parameters were varied from trial to trial, including the maximum speed of the crosshairs, the average speed of the target, the difference between crosshairs and target speeds, the number of changes in target speed (if any), the number of line segments comprising the path of each target, and the average amount of time required for the target to travel each segment.

Three scores were obtained for each trial. Two were measures of accuracy: (a) the distance from the center of the crosshairs to the center of the target at the time of firing, and (b) whether the subject "hit" or "missed" the target. The third score reflected speed and was measured by the time from trial onset until the subject fired at the target.

Multilimb Coordination

This ability does **not** apply to tasks in which trunk movement must be integrated with limb movements. It refers to tasks where the body is at rest (e.g., seated or standing) while two or more limbs are in motion.

Target Tracking Test 2. This test is very similar to the Two-Hand Coordination Test developed by the AAF. For each trial subjects are shown a path consisting entirely of vertical and horizontal lines. At the beginning of the path is a target box, and centered in the box are crosshairs. As the trial begins, the target starts to move along the path at a constant rate of speed. The subject manipulates two sliding resistors to control movement of the crosshairs. One resistor controls movement in the horizontal plane, the other in the vertical plane. The subject's task is to keep the crosshairs centered within the target at all times. This test and Target Tracking Test 1 are virtually identical except for the nature of the required control manipulation.

Number Operations

This construct involves the ability to perform, quickly and accurately, simple arithmetic operations such as addition, subtraction, multiplication, and division.

Number Memory Test. This test was modeled after a number memory test developed by Dr. Raymond Christal at the Air Force Human Resources Laboratory. Subjects are presented with a single number on the computer screen. After studying the number, the subject is instructed to push a button to receive the next part of the problem. When the button is pressed, the first part of the problem disappears and another number, along with an operation term such as Add 9 or Subtract 6 then appears. Once the subject has combined the first number with the second, he or she must press another button to receive the third part of the problem. This procedure continues until a solution to the problem is presented. The subject must then indicate whether the solution presented is right or wrong. Test items vary with respect to number of parts--four, six, or eight--contained in the single item, and the interstimulus delay period. This test is not a "pure" measure of number operations, since it also is designed to bring short-term memory into play.

Movement Judgment

Movement judgment is the ability to judge the relative speed and direction of one or more moving objects to determine where those objects will be at a given point in time and/or when those objects might intersect.

Cannon Shoot Test. The Cannon Shoot Test measures subjects' ability to fire at a moving target in such a way that the shell hits the target when the target crosses the cannon's line of fire. At the beginning of each trial, a stationary cannon appears on the video screen; the starting position varies from trial to trial. The cannon is "capable" of firing a shell, which travels at a constant speed on each trial. Shortly after the cannon appears, a circular target moves onto the screen. This target moves in a constant direction at a constant rate of speed throughout the trial, though the speed and direction vary from trial to trial. The subject's task is to push a

27

response button to fire the shell so that the shell intersects the target when the target crosses the shell's line of fire.

Three parameters determine the nature of each test trial: the angle of the target movement relative to the position of the cannon, the distance from the cannon to the impact point, and the distance from impact point to fire point.

PREDICTOR DEVELOPMENT: NON-COGNITIVE MEASURES

Two non-cognitive paper-and-pencil inventories were developed for the Pilot Trial Battery. The ABLE (Assessment of Background and Life Experiences) contains items that assess the high-priority constructs in the personality/ temperament and life history (biodata) domains. The AVOICE (Army Vocational Interest Career Examination) measures relevant constructs pertaining to vocational interests.

The extensive literature on temperament, interest, and biographical data, the results of the expert judgment study, and the covariance matrix from the preliminary battery were examined and discussed at some length in a series of meetings attended by the relevant project staff and members of the Scientific Advisory Group. The result of these deliberations was an array of constructs that were judged to be the best potential sources of valid selection/classification information of a non-cognitive nature. The linkages among the initial variable array, the constructs chosen for measurement, the variables proposed to reflect them, and the forecasted predictor/criterion correlations are shown in Figure 2-5 (Hough, 1984).

The Temperament and Biographical Measures (ABLE)

Following the identification of the construct array, item writing groups were created and items were written, revised, edited, and arranged into specific temperament and biographical scales that were intended to be valid measures of the chosen constructs. After this initial phase of item writing, revision, and scale creation, 11 substantive scales and four response bias scales were produced. Table 2-1 lists the seven constructs initially chosen for measurement via the ABLE, the 11 scales subsequently developed to represent them, and four validity scales developed under Project A. Each construct is briefly explained below.

Constructs

Adjustment. Adjustment is defined as the amount of emotional stability and stress tolerance that one possesses. The well-adjusted person is generally calm, displays an even mood, and is not overly distraught by stressful situations. He or she thinks clearly and maintains composure and rationality in situations of actual and perceived stress. The poorly adjusted person is nervous, moody, and easily irritated, tends to worry a lot, and does not do well in times of stress.

28

Figure 2-5. Linkages between literature review, expert judgments, and Preliminary and Trial Battery on Non-Cognitive Measures.

Table 2-1

Temperament/Biodata Scales (by Construct) Developed for Pilot Trial Battery:
ABLE-Assessment of Background and Life Experiences

Construct	Scale
Adjustment	Emotional Stability
Dependability	Nondelinquency Traditional Values Conscientiousness
Achievement	Work Orientation Self-Esteem
Physical Condition	Physical Condition
Leadership (Potency)	Dominance Energy Level
Locus of Control	Internal Control
Agreeableness/Likability	Cooperativeness
Response Validity Scales	Non-Random Response Unlikely Virtues (Social Desirability) Poor Impression Self-Knowledge

Dependability. The Dependability construct refers to a person's characteristic degree of conscientiousness. The dependable person is disciplined, well-organized, planful, respectful of laws and regulations, honest, trustworthy, wholesome, and accepting of authority. Such a person prefers order and thinks before acting. The less dependable person is unreliable, acts on the spur of the moment, and is rebellious and contemptuous of laws and regulations.

Achievement. Achievement is defined as the tendency to strive for competence in one's work. The achievement/work-oriented person works hard, sets high standards, tries to do a good job, endorses the work ethic, and concentrates on and persists in completion of the task at hand. This person is also confident, feels success from past undertakings, and expects to succeed in the future. The person who is less achievement-oriented has little ego involvement in his or her work, feels incapable and self-doubting, does not expend undue effort, and does not feel that hard work is desirable.

Physical Condition. The Physical Condition construct refers to one's frequency and degree of participation in sports, exercise, and physical activity.

30

Leadership (Potency). This construct was defined as the degree of impact, influence, and energy that one displays. The person high on this characteristic is appropriately forceful and persuasive, is optimistic and vital, and "gets things done." The person low on this characteristic is timid about offering opinions or providing direction and is likely to be lethargic and pessimistic.

Locus of Control. Locus of Control refers to one's characteristic belief in the amount of control one has or people have over rewards and punishments. The person with an internal locus of control expects that there are consequences associated with behavior and that people control what happens to them by what they do. Persons with an external locus of control believe that what happens is beyond their personal control.

Agreeableness/Likability. Agreeableness/Likability is defined as the degree of pleasantness versus unpleasantness exhibited in interpersonal relations. The high-scoring person is pleasant, tolerant, tactful, helpful, not defensive, and generally easy to get along with. His or her participation in a group adds cohesiveness rather than friction. The low-scoring person is critical, fault-finding, touchy, defensive, alienated, and generally contrary.

Validity Scales

The primary purpose of these scales is to determine the validity of responses, that is, the degree to which the responses are accurate depictions of the person completing the inventory.

Non-Random Response. The content (8 items) asks about information that any person is virtually certain to know.

Unlikely Virtues. This 12-item scale is aimed at detecting those who respond in a socially desirable manner (i.e., "fake good").

Poor Impression. This was an empirically derived scale designed to detect people attempting to "fake bad."

Self-Knowledge. This 13-item scale is intended to identify people who are more self-aware, more insightful, and more likely to have accurate perceptions about themselves.

The Interest Constructs/Scales (AVOICE)

The Vocational Interest Career Examination was originally developed by the Air Force. That inventory served as the starting point for the AVOICE (Army Vocational Interest Career Examination). The intent for the AVOICE was to measure all six of the constructs identified in Holland's (1966) hexagonal model of interest, as well as to provide sufficient coverage of the vocational areas most important in the Army. The six interest constructs assessed by the AVOICE, together with their associated scales, are shown in Table 2-2. The Basic Interest item, one of which is written for each Holland construct, describes a person with prototypic interests. The respondent indicates how well this description fits him or her.

31

Table 2-2

Holland Basic Interest Constructs, and Army Vocational Interest Career Examination (AVOICE) Scales Developed for Pilot Trial Battery

Construct	Scale
Realistic	Basic Interest Item
	Mechanics
	Heavy Construction
	Electronics
	Electronic Communication
	Drafting
	Law Enforcement
	Audiographics
	Agriculture
	Outdoors
	Marksman
	Infantry
	Armor/Cannon
	Vehicle Operator
	Adventure
Conventional	Basic Interest Item
	Office Administration
	Supply Administration
	Food Service
Social	Basic Interest Item
	Teaching/Counseling
Investigative	Basic Interest Item
	Medical Services
	Mathematics
	Science/Chemical
	Automated Data Processing
Enterprising	Basic Interest Item
	Leadership
Artistic	Basic Interest Item
	Aesthetics

In addition, the AVOICE included six scales dealing with organizational climate and environment and an expressed interests scale. The six constructs that pertain to a person's preference for certain types of work environments and conditions are assessed by the AVOICE through 20 scales of two items each. Figure 2-6 shows the constructs, scales, and an item from each scale.

Construct/Scale	Example
Achievement	
Achievement	"Do work that gives a feeling of accomplishment."
Authority	"Tell others what to do on the job."
Ability Utilization	"Make full use of your abilities."
Safety	
Organizational Policy	"A job in which the rules are not equal for everyone."
Supervision-Human Resources	"Have a boss that supports the workers."
Supervision-Technical	"Learn the job on your own."
Comfort	
Activity	"Work on a job that keeps a person busy."
Variety	"Do something different most days at work."
Compensation	"Earn less than others do."
Security	"A job with steady employment."
Working Conditions	"Have a pleasant place to work."
Status	
Advancement	"Be able to be promoted quickly."
Recognition	"Receive awards or compliments on the job."
Social Status	"A job that does not stand out from others."
Altruism	
Co-workers	"A job in which other employees were hard to get to know."
Moral Values	"Have a job that would not bother a person's conscience."
Social Services	"Serve others through your work."
Autonomy	
Responsibility	"Have work decision made by others."
Creativity	"Try out your own ides on the job."
Independence	"Work alone."

Figure 2-6.   AVOICE organizational climate/environment constructs, scales within constructs, and an item from each scale.

Although not a psychological construct, expressed interests were included in the AVOICE because of the extensive research indicating their validity in criterion-related studies. This Expressed Interests scale contained eight items which had three response options that formed a continuum of confidence in the person's occupational choice. Items from this scale include: "Before you went to the recruiter, how certain were you of the job you wanted in the Army?", and "If you had the opportunity right now to change your job in the Army, would you?"

As used in the pilot testing, the AVOICE included 306 items. Nearly all items were scored on a 5-point scale that ranged from "Like Very Much" (scored 5) to "Dislike Very Much" (scored 1). Items in the Expressed Interests scale were scored on a 3-point scale in which the response options were different for each item, yet one option always reflected the most interest, one moderate interest, and one the least interest.

Summary of Non-Cognitive Measures

The two non-cognitive inventories of the Pilot Trial Battery, the ABLE and the AVOICE, were designed to measure a total of 20 constructs plus a validity scale category. The ABLE assessed six temperament constructs and the Physical Condition construct through 11 scales, and also included four validity scales. Altogether, the 46 scales of the inventories included approximately 600 items.

The psychometric data obtained in pilot tests with both inventories seemed highly satisfactory; the scales were shown to be reliable and appeared to be measuring the constructs intended.

PILOT AND FIELD TESTS OF THE PILOT TRIAL PREDICTOR BATTERY

Initial Pilot Tests

Each instrument in each category (cognitive paper-and-pencil, computerized, and non-cognitive) was pilot tested one or more times with various small samples from Fort Campbell, Fort Carson, and Fort Lewis. Based on feedback from the respondents, refinements were made in directions, format, and item wording. A few items were dropped because of extreme item statistics. However, the basic structure of each instrument remained the same until more data from the larger scale field tests became available.

Field Tests

The final step before the Concurrent Validation was a more systematic series of field tests of all the predictor measures, using larger samples. The outcome of the field test/revision process was the final form of the predictor battery (i.e., the Trial Battery) to be used in the Concurrent Validation.

Field tests were conducted at three sites. The sites and basic purposes of the field test at each site are described below.

Fort Knox - The full Pilot Trial Battery (PTB) was administered here to evaluate the psychometric characteristics of all the measures and to analyze the covariance of the measures with each other and with the ASVAB. In

34

addition, the measures were readministered to part of the sample to provide data for estimating the test-retest reliability of the measures. Finally, part of the sample received practice on some of the computer measures and were then retested to obtain an estimate of the effects of practice on scores on computer measures.

Fort Bragg - The non-cognitive Pilot Trial Battery measures, Assessment of Background and Life Experiences (ABLE) and Army Vocational Interest Career Examination (AVOICE), were administered to soldiers at Fort Bragg under several experimental conditions to estimate the extent to which scores on these instruments could be altered or "faked" when persons are instructed to do so.

Minneapolis Military Entrance Processing Station - The non-cognitive measures were administered to a sample of recruits as they were being processed into the Army, to obtain an estimate of how persons in an applicant setting might alter their scores.

Results of the Cognitive Paper-and-Pencil and Computer-Administered Field Tests

Psychometric Data

The basic data obtained on the cognitive paper-and-pencil and the computer-administered tests are portrayed in Tables 2-3 and 2-4, respectively.

Factor Analysis Results

Two variables, PS&A reaction time and Short-Term Memory reaction time, were omitted because the reaction time scores from these measures correlated very highly with their corresponding slope or intercept variables. Results from the seven-factor solution of a principal components factor analysis with varimax rotation are displayed in Table 2-5. All loadings of .30 or greater are shown.

Factor 1 includes eight of the ASVAB subtests, six of the paper-and-pencil measures, and two cognitive/perceptual computer variables. Because this factor contains measures of verbal, numerical, and reasoning ability, it was termed "g", to represent general cognitive ability.

Factor 2 was a general spatial factor and included all of the PTB cognitive paper-and-pencil measures, Mechanical Comprehension from the ASVAB, and Target Identification reaction time from the computer tests.

Factor 3 loaded on the three psychomotor tests, with substantially smaller loadings from three cognitive/perceptual computer test variables, the Path Test, and Mechanical Comprehension from the ASVAB. Given the high loadings of the psychomotor tests, it was labeled the motor factor.

Factor 4 included variables from the cognitive/perceptual computer tests. This factor appears to involve accuracy of perception across several tasks and types of stimuli.

For Factor 5, the highest loadings were on straightforward reaction time measures. Consequently, it was interpreted as a speed of reaction factor.

35

Table 2-3

Means, Standard Deviations, and Reliability Estimates for the Ten Paper-and-Pencil Cognitive Tests; Fort Knox Field Tests

Test	No. of Items	Time Allotted (in minutes)	Score Mean[a]	SD[c]	Reliability Coefficient[b]		
					Split Half (N = 118)	Coefficient Alpha (N = 290)	Test-Retest (N = 97 to 126)
Assembling Objects	40	16	26.5	8.7	.79	.92	.74
Object Rotation	90	7.5	59.6	19.0	.86	.97	.75
Path	44	8	26.4	10.2	.82	.92	.64
Maze	24	5.5	17.8	4.5	.78	.89	.71
Shapes	54	16	25.4	8.9	.82	.92	.70
Orientation 1	150	10	19.6	5.2	.92	.98	.67
Orientation 2	24	10	21.6	3.6	.89	.88	.80
Orientation 3	20	12	88.7	34.8	.88	.90	.84
Reasoning 1	30	12	11.5	6.0	.78	.83	.64
Reasoning 2	32	10	7.7	5.7	.63	.65	.57

[a]Ns range from 292 to 298 for mean and SD calculations.

[b]The split-half coefficient is computed on pilot test data from Fort Lewis, where two separately timed halves were given, and is corrected to full test length. Coefficient alphas are based on the Fort Knox data and are overestimates for the speeded tests. The test-retest interval was two weeks.

35

Table 2-4

Characteristics of the 19 Dependent Measures for Computer-Administered Tests:
Fort Knox Field Tests

Dependent Measure	Mean[a]	SD[a]	Split-Half (r_{sh})[b]	Test-Retest (r_{tt})[b]
PERCEPTUAL				
Simple Reaction Time (SRT)	56.2 hs[c]	18.8 hs	.90	.37
Mean Reaction Time (RT)				
Choice Reaction Time (CRT)				
Mean Reaction Time (RT)	67.4 hs	10.2 hs	.89	.56
Perceptual Speed and Accuracy (PS&A)				
Percent Correct (PC)	88%	8%	.83	.59
Mean Reaction Time (RT)	325.6 hs	70.4 hs	.96	.65
Slope	42.7 hs/ch[d]	15.6 hs/ch	.88	.67
Intercept	68.0 hs	45.0 hs	.74	.55
Target Identification				
Percent Correct (PC)	90%	10%	.84	.19
Mean Reaction Time (RT)	528.7 hs	134.0 hs	.96	.67
Short-Term Memory (STM)				
Percent Correct (PC)	85%	8%	.72	.34
Mean Reaction Time (RT)	129.7 hs	23.8 hs	.94	.78
Slope	7.2 hs/ch	4.5 hs/ch	.52	.47
Intercept	108.1 hs	23.2 hs	.84	.74
Number Memory				
Percent Correct (PC)	83%	13%	.63	.53
Mean Operation Time (RT)	230.7 hs	73.9 hs	.95	.88
Cannon Shoot				
Time Error (TE)	78.6 hs	20.3 hs	.88	.66
PSYCHOMOTOR				
Target Track 1				
Mean Log Distance	3.2	.44	.97	.68
Target Shoot				
Mean Time to Fire (std) (TF)	-.01	.48	.91	.48
Mean Log Distance (std)	-.01	.41	.86	.58
Target Track 2				
Mean Log Distance	3.91	.49	.97	.68

[a] N = 256, but varies slightly from test to test.
[b] N = 120 for test-retest reliabilities, but varies slightly from test to
   test. r_{sh} = split-half reliability; odd-even item correlation with
   Spearman-Brown correction. r_{tt} = test-retest reliability, 2-week
   interval between administrations.
[c] hs = hundredths of a second.

Table 2-5

Principal Components Factor Analysis of Scores of the ASVAB Subtests, Cognitive Paper-and-Pencil Measures, and Perceptual/Psychomotor Computer-Administered Tests[a] (N ≈ 169)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	h²
ASVAB								
GS	75							59
AR	75							73
WK	77							62
PC	62							47
NO						84		77
CS						62		44
AS	62							58
MK	77							70
MC	63	38	-30					68
EI	72							65
COGNITIVE PAPER-AND-PENCIL								
Assemb Obj	35	69						66
Obj Rotation		-61						49
Shapes		66						51
Maze		70						67
Path		67	-30					65
Reason 1	37	58						54
Reason 2	37	47						44
Orient 1	37	64						58
Orient 2	40	46			-30			52
Orient 3	60	52						67
PERCEPTUAL COMPUTER								
SRT-RT					63			44
CRT-RT					61			50
PS&A-PC				67	31			70
PS&A Slope				88				81
PS&A Inter				-65	50			74
Target ID-PC				40				25
Target ID-RT		-41	37		30			57
STM-PC				39			34	41
STM-Slope							41	25
STM-Int			38		51			47
Cannon Shoot-TE			32					19
No Mem-PC	53					37		52
No Mem-RT	-37					-46		54
PSYCHOMOTOR COMPUTER								
Tracking 1			86					82
Tracking 2			77					66
Target Shoot-TF						42		23
Target Shoot-Dist			64					48
Eigen Values	5.69	4.70	2.83	2.37	1.92	1.87	1.17	

Note: Decimals have been omitted from factor loadings.

[a]Note that the following variables were not included in this factor analysis: AFQT, PS&A Reaction Time, and Short-Term Memory Reaction Time.

h² - communality (sum of squared factor loadings) for variables.

Factor 6 contained four variables, two from the ASVAB and two from the cognitive/perceptual computer tests. This factor appears to represent both speed of reaction and arithmetic ability.

Factor 7 contains three variables from the computer tests. These include Short-Term Memory percent correct and slope, and Target Shoot time-to-fire. This factor is difficult to interpret, but was believed to represent a response style factor. That is, this factor suggests that those individuals who take a longer time to fire on the Target Shoot Test also tend to have higher slopes on the Short-Term Memory (lower processing speeds with increased bits of information) but are more accurate or obtain higher percent correct values on the Short-Term Memory Test.

Note that several variables have fairly low commonalities. These may be due to relatively low score variance or reliability, but could also be due to those variables having unique variance, at least when factor analyzed with this set of tests. This latter explanation was seen as more highly plausible for the Cannon Shoot score.

Special Analyses on Computer-Administered Tests

Correlations with Video Game-Playing Experience. Field test subjects were asked the question, "In the last couple years, how much have you played video games?" The five possible alternatives ranged from "You have never played video games" to "You have played video games almost every day" and were given scores of 1 to 5, respectively. The mean was 2.99, SD was 1.03 (N = 256), and the test-retest reliability was .71 (N = 113).

The 19 correlations of this item with the computer test scores ranged from -.01 to +.27, with a mean of .10. A correlation of .12 is significant at alpha = .05. These findings were interpreted as showing a small, but significant, relationship of video game-playing experience to the more "game-like" tests in the battery.

Practice Effects on Selected Computer Test Scores. The results of the analyses of variance for the five tests included in the practice effects research (Table 2-6) show only one statistically significant practice effect, the Mean Log Distance score on Target Tracking Test 2. There were three statistically significant findings for time, indicating that scores did change with a second testing, whether or not practice trials intervened between the two tests. Finally, the Omega squared value indicates that relatively small amounts of test score variance are accounted for by the Group, Time, or Time by Group factors.

These data suggest that the practice intervention was not a particularly strong one. The average gain score for the two groups across the five dependent measures was only .09 standard deviation.

Table 2-6

Effects of Practice on Selected Computer Test Scores

Test	Dependent Measure	Source of Variance	df	F	Omega Squared
Choice Reaction Time	Trimmed Mean Reaction Time	Group	1,180	9.71*	.032
		Time	1,180	25.70*	.035
		Time x Group	1,180	.73	---
Target Tracking 1	Mean Log Distance	Group	1,178	.73	---
		Time	1,178	9.26*	.005
		Time x Group	1,178	4.11	---
Target Tracking 2	Mean Log Distance	Group	1,178	.47	---
		Time	1,178	1.30	---
		Time x Group	1,178	7.79*	.005
Cannon Shoot	Time Error	Group	1,171	3.79	---
		Time	1,171	.16	---
		Time x Group	1,171	5.72	---
Target Shoot	Mean Log Distance	Group	1,171	.41	---
		Time	1,171	9.28*	.012
		Time x Group	1,171	.08	---

*Denotes significance at $p < .01$.

Field Test Results for the Non-Cognitive Measures (ABLE and AVOICE)

Psychometric Data

The Fort Knox data were used to obtain descriptive scale statistics and examine the covariation among scales. Summary statistics for the ABLE and AVOICE are presented in Tables 2-7 (ABLE) and 2-8 (AVOICE). The median alpha coefficient (internal consistency) for the ABLE content scales is .84, and the median test-retest (2-week interval) correlation is .79, with a range of .68 to .83. The median alpha coefficient for the AVOICE scales is .86, and the median test-retest correlation is .76.

Fakability Analyses

To investigate intentional distortion of responses, data were gathered (a) from soldiers instructed, at different times, to distort their responses or to be honest (experimental data gathered at Fort Bragg); (b) from soldiers who were simply responding to the ABLE and AVOICE with no particular directions (data gathered at Fort Knox); and (c) from recently sworn-in Army recruits at the Minneapolis MEPS.

40

Table 2-7

ABLE Scale Score Characteristics: Fort Knox Field Test (N = 276 except where otherwise noted)

Scale	No. of Items	Mean Time 1	SD	Alpha	Test-Retest[a] r	Median Item-Scale[b] r
Content Scales						
Emotional Stability	29	64.9	8.27	.86	.68	.44
Self-Esteem	15	35.1	5.25	.83	.81	.54
Cooperativeness	24	54.1	6.09	.77	.69	.42
Conscientiousness	21	48.9	5.90	.81	.73	.43
Nondelinquency	24	55.4	7.23	.84	.81	.46
Traditional Values	16	37.2	4.60	.70	.74	.45
Work Orientation	27	61.2	7.93	.85	.80	.47
Internal Control	21	50.3	6.14	.79	.75	.43
Energy Level	25	57.1	7.11	.85	.79	.47
Dominance	16	35.5	6.13	.86	.83	.56
Physical Condition	9	31.1	7.53	.87	.81	.72
Response Validity Scales						
Unlikely Virtues	12	16.6	3.39	.68	.62	.53
Self-Knowledge	13	29.6	3.54	.62	.71	.41
Non-Random Response[c]	8	7.7	.71	.56	.37	.45
Poor Impression	24	1.5	1.86	.61	.56	.33

[a]N=109 for test-retest correlations. Test-retest interval was 2 weeks.

[b]Without item deletion.

[c]N=281. Statistics reported for this scale are based on sample edited for overall missing data only. "Passing" score on non-random response scale $\leq$ 6.

Table 2-8

AVOICE Scale Score Characteristics: Fort Knox Field Test (N = 270 except where otherwise noted)

Scale	No. of Items	Mean	SD	Alpha	Test-Retest[a] r	Median Item-Scale r
Marksman	5	15.8	4.37	.79	.77	.75
Agriculture	5	14.1	3.99	.68	.65	.70
Mathematics	5	15.1	4.37	.82	.76	.79
Aesthetics	5	14.3	4.17	.77	.72	.74
Leadership	6	20.3	4.70	.81	.55	.74
Electronic Communication	7	21.1	5.73	.92	.78	.72
Automated Data Processing	7	23.4	6.56	.88	.81	.81
Teaching/Counseling	7	22.8	5.53	.82	.73	.73
Drafting	7	21.5	6.12	.85	.74	.77
Audiographics	7	23.8	5.68	.82	.76	.70
Armor/Cannon	8	22.4	6.57	.83	.74	.69
Vehicle/Equipment Operator	10	28.1	7.79	.86	.69	.70
Outdoors	9	31.7	6.41	.79	.65	.66
Infantry	10	29.1	7.13	.81	.78	.65
Science/Chemical Operations	11	29.4	8.93	.89	.79	.71
Supply Administration	13	35.0	10.44	.92	.82	.75
Office Administration	16	45.2	13.20	.94	.86	.73
Law Enforcement	16	48.1	11.64	.88	.78	.63
Mechanics	16	50.0	14.68	.95	.80	.80
Electronics	20	60.0	17.48	.96	.74	.77
Heavy Construction/Combat	23	65.8	17.90	.94	.76	.70
Medical Services	24	68.5	18.79	.95	.84	.69
Food Service	17	48.2	11.16	.89	.71	.64

[a] n=127 for test-retest correlations. Test-retest interval was 2 weeks.

The purposes of the faking study were to:

- Determine the extent to which soldiers can distort their responses to temperament and interest inventories when instructed to do so. (Compare data from Fort Bragg faking conditions with Fort Bragg and Fort Knox honest conditions.)

- Determine the extent to which the ABLE response validity scales detect such intentional distortion. (Compare response validity scales in Fort Bragg honest and faking conditions.)

- Determine the extent to which ABLE validity scales can be used to correct or adjust scores for intentional distortion.

- Determine the extent to which distortion is a problem in an applicant setting. (Compare MEPS data with Fort Bragg and Fort Knox data.)

The participants in the experimental group were 425 enlisted soldiers in the 82nd Airborne Brigade at Fort Bragg. Comparison samples were MEPS candidates (N = 126) and the Fort Knox soldiers described earlier (N = 276).

Four faking and two honest conditions were created:

ABLE - Fake Good

 Imagine you are at the Military Entrance Processing
 Station (MEPS) and you want to join the Army.
 Describe yourself in a way that you think will ensure
 that the Army selects you.

ABLE - Fake Bad

 Imagine you are at the Military Entrance Processing
 Station (MEPS) and you do not want to join the Army.
 Describe yourself in a way that you think will ensure
 that the Army does not select you.

ABLE - Honest

 You are to describe yourself as you really are.

AVOICE - Fake Combat

 Imagine you are at the Military Entrance Processing
 Station (MEPS). Please describe yourself in a way
 that you think will ensure that you are placed in an
 occupation in which you are likely to be exposed to
 combat during a wartime situation.

AVOICE - Fake Non-combat

 Imagine you are at the Military Entrance Processing
 Station (MEPS). Please describe yourself in a way you
 think will ensure that you are placed in an occupation

43

in which you are <u>unlikely</u> to be exposed to combat during a wartime situation.

AVOICE - Honest

You are to describe yourself as you really are.

The design was a 2x2x2 with repeated measures on faking and honest conditions which were counterbalanced. Thus, approximately half the experimental group, 124 soldiers, completed the inventories honestly in the morning and faked in the afternoon, while the other half (121) completed the inventories honestly in the afternoon and faked in the morning. The first between-subjects factor consisted of these two levels: Fake Good/Want Combat and Fake Bad/Do Not Want Combat. Order was manipulated in the second between-subjects factor such that the following two levels were produced: Faked responses then honest responses, and honest responses then faked responses.

Overall, the ABLE data supported the following conclusions:

- Soldiers can distort their responses when instructed to do so.

- The response validity scales detect intentional faking.

- An individual's Social Desirability scale score can be used to adjust his or her content scale scores to reduce variance associated with faking.

- Faking or distortion may not be a significant problem in an applicant setting.

Overall, the AVOICE data showed the following:

- Soldiers can distort their responses when instructed to do so.

- The ABLE Social Desirability and Poor Impression scales are not as effective for adjusting AVOICE scale scores as they are for adjusting ABLE content scale scores.

- Faking or distortion may not be a significant problem in an applicant setting.

TRANSFORMING THE PILOT TRIAL BATTERY INTO THE TRIAL BATTERY

In the field tests the entire Pilot Trial Battery required approximately 6.5 hours of administration time. However, the Trial Battery, which was the label reserved for the predictor battery to be used in the full-scale Concurrent Validation, had to fit in a 4-hour time slot.

Using all the accumulated information, final revisions were made during a series of meetings attended by the project staff and by the Scientific Advisory Group. The revisions and the stated reasons for their adoption are summarized below.

44

Changes to Cognitive Paper-and-Pencil Tests

Changes to the cognitive paper-and-pencil tests are summarized in Table 2-9.

The Spatial Visualization construct was measured by three tests: Assembling Objects, Object Rotation, and Shapes. The Shapes Test was dropped because the previous evidence of validity for predicting job performance was judged to be less impressive than for the other two tests. Eight items were dropped from the Assembling Objects Test by eliminating items that were very difficult or very easy, or had low item-total correlations. The time limit was not changed, which made it more a power test than before.

For the Spatial Scanning construct, the Path Test was dropped and the Mazes Test was retained with no changes. Mazes was a shorter test, showed higher test-retest reliabilities (.71 vs. .64), and gain scores were lower (.24 vs. .62 SD unit).

Table 2-9

Summary of Changes to Paper-and-Pencil Cognitive Measures in the Pilot Trial Battery

Test Name	Changes
Assembling Objects	Decrease from 40 to 32 items.
Object Rotation	Retain as is with 90 items.
Shapes	Drop test.
Mazes	Retain as is with 24 items.
Path	Drop Test.
Reasoning 1	Retain as is with 30 items. New name: REASONING TEST.
Reasoning 2	Drop Test.
Orientation 1	Drop Test.
Orientation 2	Retain as is with 24 items. New name: ORIENTATION TEST.
Orientation 3	Retain as is with 20 items. New name: MAP TEST.

Reasoning Test 1 was evaluated as the better of the two tests for Figural Reasoning because it had higher reliabilities as well as a higher uniqueness estimate. It was retained with no item or time limit changes, and Reasoning Test 2 was dropped.

Of the three tests that measured the Spatial Orientation construct, Orientation Test 1 was dropped because it showed lower test-retest reliabilities (.67 vs. .80 and .84) and higher gain scores (.63 SD unit vs. .11 and .08 SD unit).

Changes to Computer-Administered Tests

Besides the changes made to specific tests, several improvements were made to the computer battery as a whole. The general changes designed to save time were as follows:

- Most instructions were shortened considerably.

- Whenever the practice items had a correct response, the subject was given feedback.

- Rest periods were eliminated. This was possible because virtually every test was shortened.

- The total time allowed for subjects to respond to a test item (i.e., response time limit) was set at 9.0 seconds for all reaction time tests.

Changes to the individual computer-administered tests are summarized in Table 2-10.

Fifteen items were added to Choice Reaction Time in an attempt to increase the test-retest reliability for mean reaction time.

Twelve items were eliminated from the Perceptual Speed and Accuracy Test (reduced from 48 to 36 items), primarily to save time. Reduction in the number of items did not seem to be cause for reliability concerns.

Several changes were made to the Target Identification Test. First, the "moving" items were eliminated; field test data showed that scores on the "moving" and stationary items correlated .78, and the moving items had lower test-retest reliabilities than stationary items (.54 vs. .74). All target objects were made the same size since field test analyses indicated size had no appreciable effect on reaction time. A third level of angular rotation was added so that the target objects were rotated either 0°, 45°, or 75°. Finally, the number of items was reduced from 48 to 36 to save time; internal consistency and test-retest estimates indicated that the level of risk attached to this reduction was acceptable.

Analyses of field test data showed the probe delay period difference did not significantly affect mean reaction time scores, so it was eliminated from the Short-Term Memory Test. To save time, 12 items were eliminated.

46

Table 2-10

Summary of Changes to Computer-Administered Measures in the
Pilot Trial Battery

Test Name	Changes
COGNITIVE/PERCEPTUAL TESTS	
Demographics	Eliminate race, age, and typing experience items. Retain SSN and video experience items.
Simple Reaction Time	No changes.
Choice Reaction Time	Increase number of items from 15 to 30.
Perceptual Speed	Reduce items from 48 to 36. Eliminate word & Accuracy items.
Target Identification	Reduce items from 48 to 36. Eliminate moving items. Allow stimuli to appear at more angles of rotation.
Short-Term Memory	Reduce items from 48 to 36. Establish a single item presentation and probe delay period.
Cannon Shoot	Reduce items from 48 to 36.
Number Memory	Reduce items from 27 to 18. Shorten item strings. Eliminate item part delay periods.
PSYCHOMOTOR TESTS	
Target Tracking 1	Reduce items from 27 to 18. Increase item difficulty.
Target Tracking 2	Reduce items from 27 to 18. Increase item difficulty.
Target Shoot	Reduce items from 40 to 30 by eliminating the extremely easy and extremely difficult items.

The number of items on the Cannon Shoot Test was reduced from 48 to 36. Reliabilities for the time error scores were high enough to warrant such reduction without the expectation of a significant impact on reliability.

Two modifications were made to Number Memory to reduce test administration time. The item delay period was made a constant (1 second) rather than treated as a parameter with two levels (0.5 and 2.5 seconds), and the item string length (number of parts in an item) was changed from 4, 6, or 8 parts to 2, 3, or 4 parts. These changes drastically reduced the time required to complete the test.

Similar kinds of changes were made to Target Tracking Tests 1 and 2. Since internal consistency and test-retest reliability estimates were relatively high, the number of items was reduced from 27 to 18.

Several changes were made to the Target Shoot Test. First, all test items were classified according to three parameters: crosshairs speed, ratio of target to crosshairs speed, and item complexity (i.e., number of turns/mean segment length). Then, items were revised to achieve a balanced number of items in each cell when the levels of these parameters were crossed. Second, extremely difficult items were eliminated and item presentation times (the time the target was visible on the screen) were increased to a minimum of 6 seconds (and a maximum of 10 seconds). This was done to eliminate a severe missing data problem for such items which seemed to occur when the target moved very rapidly, made many sudden changes in direction and speed, or was shown only a few seconds. The number of items was reduced from 40 to 30 to save testing time.

Changes to Non-Cognitive Measures (ABLE and AVOICE)

Changes to the non-cognitive measures (ABLE and AVOICE) are summarized in Table 2-11. Time constraints required a 25 percent reduction in the total number of ABLE and AVOICE items. The goal was to decrease items on a scale-by-scale basis, while preserving the basic content of each scale. A decision was also made to delete the Agriculture scale, the six single-item Holland scales, and the eight Expressed Interest items.

Table 2-11

Summary of Changes to Pilot Trial Battery Versions of Assessment of
Background and Life Experiences (ABLE) and Army Vocational Interest
Career Examination (AVOICE)

Inventory/Scale Name	Changes
ABLE-Total	Decrease from 270 to approximately 209 items.
AVOICE-Total	Decrease from 309 to approximately 214 items.
AVOICE Expressed Interest Scales	Drop.
AVOICE Single Item Holland Scales	Drop.
AVOICE Agriculture Scale	Drop.
Work Environment Preference Scales	Move to criterion measure booklet (Delete from AVOICE booklet).

The Trial Battery

The final array of tests for the Trial Battery is shown in Table 2-12.
The Trial Battery was designed to be administered in a period of 4 hours
during the Concurrent Validation phase of Project A, in which data collection
began in FY85. Data collected in that phase would allow the first look at the
validity of Trial Battery measures against job performance criteria.

49

Table 2-12

Description of Measures in the Trial Battery

Cognitive Paper-and-Pencil Tests	Number of Items	Time Limit (minutes)
Reasoning Test	30	12
Object Rotation Test	90	7.5
Orientation Test	24	10
Maze Test	24	5.5
Map Test	20	12
Assembling Objects Test	32	16

Computer-Administered Tests	Number of Items	Approximate Time
Demographics	2	4
Reaction Time 1	15	2
Reaction Time 2	30	3
Memory Test	36	7
Target Tracking Test 1	18	8
Perceptual Speed and Accuracy Test	36	6
Target Tracking Test 2	18	7
Number Memory Test	28	10
Cannon Shoot Test	35	7
Target Identification Test	36	4
Target Shoot Test	30	5

Non-Cognitive Paper-and-Pencil Inventories	Number of Items	Approximate Time
Assessment of Background and Life Experiences (ABLE)	209	35
Army Vocational Interest Career Examination (AVOICE)	176	20

Chapter 3
CRITERION DEVELOPMENT

INTRODUCTION

The overall goals of measuring training and job performance--that is, criteria--in Project A were to define the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor. The specific measures were used as criteria against which to validate selection and classification tests and were not at the outset intended to serve as operational methods for appraising performance. The research participants were informed that the measures would not be entered into their personnel file.

The Developmental Approach

The general procedure for criterion development in Project A followed a basic cycle of a comprehensive literature review, conceptual development, scale construction, pilot testing, scale revision, field testing, and proponent (management) review. The specific measurement goals were to:

- Make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency.

- Compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e., a multitrait, multimethod approach).

- Develop rating scale measures of performance factors that are common to all first-tour enlisted MOS (Army-wide measures), as well as for factors that are specific to each MOS.

- Develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance.

- Exploit existing file/administrative data as much as possible for indicators of individual performance.

- Use the data from the Concurrent Validation sample to develop a model of the latent structure of job performance in first-tour enlisted MOS.

Given these intentions, the criterion development effort focused on three major methods of measuring performance: hands-on job sample tests, multiple-choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in developing the rating scales.

The Modeling of Performance

The criterion development efforts to be described were guided by a particular theory of performance. The intent was to proceed through an almost continual process of data collection, expert review, and model/theory revision.

Multidimensionality

As a basic concept, job performance was viewed as multidimensional. There is not one attribute, one outcome, one factor, or one anything that can be pointed to and labeled as job performance. Further, job performance was given the status of a construct (which implies a "theory" of performance), and is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization. For example, a manager could make contributions to organizational goals by working out congruent short-term goals for his subordinates, and thereby guiding them in the right direction, or by praising them for a job well done, and thereby increasing subsequent effort levels. Each of these activities probably requires different knowledges and skills, which are in turn most likely a function of different abilities.

Consequently, for any particular job, one fundamental task of performance measurement is to describe the basic factors that comprise performance. That is, how many such factors are there and what is their basic nature?

Two General Types of Factors

For the population of entry-level enlisted positions in the Army, there should be two major types of job performance factors: components that reflect MOS-specific technical competence or specific job behaviors that are not required for other jobs, and components that are defined and measured in the same way for every job. The latter have been referred to as "Army-wide" criterion factors, such as performance on the common tasks for which every soldier is responsible.

The Army-wide concept incorporates the basic notion that total performance is much more than task or technical proficiency. It might include such things as contribution to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance in the face of adversity. A much more detailed description of the initial working model for the Army-wide segment of performance can be found in Borman, Motowidlo, Rose, and Hanser (1986).

In sum, the working model of total performance with which the project began viewed performance as multidimensional within the two broad categories of factors or constructs. The job analysis and criterion construction methods were designed to describe the content of these factors via an extensive description of the total performance domain, several iterations of data collections, and the use of multiple methods for identifying basic performance factors.

52

Factors Versus a Composite

Saying that performance is multidimensional does not preclude using just one index of an individual's contributions to make a specific personnel decision (e.g., select/not select, promote/not promote). As argued by Schmidt and Kaplan (1971) some years ago, it seems quite reasonable for the organiza-tion to scale the importance of each major performance factor relative to a particular personnel decision that must be made, and to combine the weighted factor scores into a composite that represents the total contribution or utility of an individual's performance, within the context of that decision. That is, the way in which performance information is weighted is a value judgment on the organization's part. The deter-mination of the specific combinational rules (e.g., simple sum, weighted sum, nonlinear combination) that best reflect what the organization is trying to accomplish is in large measure a research question.

A Structural Model

If performance is characterized in the above manner, then a more formal way to model performance is to think in terms of its latent structure. The usual common factor model of the latent structure is open to criticism because all of the criterion (i.e., performance) measures may not be at the same level of explanation, or they may be so qualitatively different that putting them into the same correlation matrix does not seem appropriate, or two criteria may not be functionally independent. One might be a cause of another; for example, individual differences in training performance may be a cause of individual differences in job performance.

From this perspective, the aims of criterion analysis are to use all available evidence, theory, and professional judgment to (a) identify the variables that are necessary and sufficient to explain the phenomena of interest, and (b) specify the nature of the relationships between pairs of variables in terms of whether they 1) are correlated because one is a cause of another, 2) are correlated because both are manifestations of the same latent property, or 3) are independent. The more explicitly the causal directions and the predicted magnitude of the associations can be specified, the greater the potential power of the model if it is confirmed by subsequent empirical data.

Within the structural equation framework there are manifest variables (operational measures) and latent variables (constructs). The Project A proposal and research plan dealt explicitly with criterion constructs and criterion measures.

A few points should be made about this view. First, a lot more is known about predictor (i.e., ability, temperament, and interest) constructs than about job performance constructs. There are volumes of research on the former, and almost none on the latter. Relatively little attention has been given to conceptualizing performance in clerical, technical, or skilled jobs.

Second, the usual textbook illustration of a latent structural equation model (e.g., James, Muliak, & Brett, 1982) shows each latent variable being represented by one or more manifest operational measures. However, just as it is easy to think of examples where a predictor test score could be a function of more than one latent variable (e.g., the score on computerized two-hand

53

tracking apparatus could be a function of several latent psychomotor "factors"), the same will be true of criterion measures. Most of them will not be unidimensional.

Third, we would be hard-pressed to defend placing the criterion variables on some continuum from immediate to intermediate to ultimate as a means for portraying their relative importance or their functional interrelationships.

Finally, people do not usually work alone. Individuals are members of work groups or units and it is the unit's performance that frequently is the most central concern of the organization. However, determining the individual's contribution to the unit's score is not a simple problem. Further, variation in unit performance is most likely a function of a number of factors besides the "true" level of performance of each individual. The quality of leadership, weather conditions, or the availability of spare parts are examples of such additional sources of variation in unit performance.

In sum, Project A researchers attempted, in state-of-the-art fashion, to develop both a theory of entry-level performance in skilled jobs (i.e., as represented by the population of Army MOS) and to construct multiple valid and reliable measures of each major performance component. In large measure, the project was successful in doing so and has now gone far beyond any previous efforts to account for the totality of job performance.

CRITERION DEVELOPMENT: MOS-SPECIFIC TASK-BASED PERFORMANCE MEASURES

The task analysis-based, MOS-specific criterion measures concern the assessment of performance on a sample of tasks for a particular MOS that were identified as representative of all tasks in that MOS. The general procedure was to develop a careful description of all the major tasks that comprise the job (i.e., the total population or domain of tasks), draw a sample of these tasks, and develop multiple measures of performance on each task. (See Campbell, C. H., Campbell, R. C., Rumsey, & Edwards, 1986.)

The total number of tasks to be sampled was dictated primarily by time constraints. While the time required to assess performance on individual tasks would differ by task, a total of 30 tasks for each MOS was taken as a reasonable planning figure.

For each MOS, all 30 tasks would be assessed with written knowledge tests. Fifteen of the 30 tests would also be assessed with hands-on tests. Finally, task performance ratings would be obtained for the 15 tasks measured with the hands-on job sample tests, and job history data covering recency and frequency of performance would be obtained for all 30 tasks. As noted previously, because of cost considerations the MOS-specific job performance measures (i.e., the hands-on tests and MOS-specific ratings) could be developed for only nine of the 19 original MOS in the sample. The nine were further divided into two groups known as Batch A and Batch B. The MOS in Batch A were done first; sometimes during the development period the lessons learned in Batch A led to changes in procedures for Batch B. The remaining 10 MOS became known as Batch Z. The compositions of Batches A, B, and Z are shown in Table 3-1.

Table 3-1

MOS Grouping for Criterion Development

	MOS
Batch A	13B Cannon Crewman
	64C Motor Transport Operator
	71L Administrative Specialist
	95B Military Police
Batch B	11B Infantryman
	19E Armor Crewman
	31C Radio Teletype Operator
	63B Light Wheel Vehicle Mechanic
	91A Medical Specialist
Batch Z	12B Combat Engineer
	16S MANPADS Crewman
	27E TOW/Dragon Repairer
	51B Carpentry/Masonry Specialist
	54E Chemical Operations Specialist
	55B Ammunition Specialist
	67N Utility Helicopter Repairer
	76W Petroleum Supply Specialist
	76Y Unit Supply Specialist
	94B Food Service Specialist

Defining the Task Domain

Enumerating the total task domain for an MOS was based on three primary sources:

MOS-Specific Soldier's Manuals (SM). Each MOS Proponent, the agency responsible for prescribing MOS policy and doctrine, prepares and publishes a Soldier's Manual that lists and describes tasks, by skill level, that soldiers in the MOS are doctrinally responsible for knowing and performing. The number of tasks per MOS varies widely from a low of 17 Skill Level 1 (SL1) tasks to more than 130 SL1 tasks.

Soldier's Manual of Common Tasks (SMCT). The SMCT describes tasks that each soldier in the Army, regardless of his or her MOS, must be able to perform. The 1983 version contained 78 SL1 tasks.

Army Occupational Survey Program (AOSP). The AOSP obtains task descriptions by surveying job incumbents with a questionnaire checklist that includes several hundred items. The items are obtained from a variety of sources (e.g., the Proponent school), and include and expand the doctrinal tasks from the preceding two sources. The AOSP is administered to soldiers in all skill levels of each MOS by the U.S. Army Soldier Support Center. The analysis of responses by means of the Comprehensive Occupational Data Analysis Program (CODAP) provides the number and percentage of soldiers at each skill level who report that they perform each task. The number of activities in the AOSP surveys for the nine MOS of interest ranged from 546 to well over 800.

Proponent agencies were also contacted directly to determine whether relevant tasks existed beyond those listed in the three primary sources. The number of additional tasks thus generated was not large, but the tasks were sometimes significant. For example, the introduction of new equipment added tasks that had not yet appeared in the written documentation.

The preliminary aggregate list of SM/SMCT tasks and AOSP statements was carefully edited for redundancies, and items were revised and combined to achieve a relatively uniform level of generality and format across items. The result was a refined list of MOS tasks used as a basis for domain review and consolidation.

At each Proponent a minimum of three senior NCOs or officers reviewed the refined list for an MOS. These subject matter experts (SME) eliminated tasks that had been incorrectly included in the domain, for reasons such as equipment that was being changed, current doctrine not yet reflected in available publications, and equipment variations that should be combined.

In the final phase, the task lists resulting from domain consolidation were again reviewed to eliminate tasks that pertained to restricted duty positions or that were performed only infrequently. The result of this process was a final task list (or population) for each MOS. Table 3-2 shows the reduction of the task list during each phase and the reasons for the reduction, by MOS.

SME Judgments of Task Characteristics

As preparation for selecting 30 representative tasks for each MOS, 15-30 SMEs (NCOs at E6 or above and officers at grade O-3 or above) rated each task on a number of characteristics. Three types of judgments were obtained:

Task Clustering. Each task was listed on a 3 x 5 inch card along with a brief description. SMEs were told to sort the tasks into groups so that all the tasks in each group were alike, and each group was different from the other groups.

Task Importance. The procedure for rating task importance was different for the first four MOS (Batch A) than for the last five MOS (Batch B) that were analyzed (see Table 3-1). For Batch A, all SMEs were given a European scenario that specified a high state of training and strategic readiness but was short of involving actual conflict. After Batch A data were collected, concern was expressed as to the scenario effect on SME judgments. As a result, for Batch B three scenarios were used. An "Increasing Tension"

Table 3-2
Effects of Domain Definition on MOS Task Lists

	MOS								
	13B	**64C**	**71L**	**95B**	**11B**	**19E**	**31C**	**63B**	**91A**

AUSP Review									
AOSP Statements	669	677	822	546	822	609	656	633	685
Deleted - Zero Frequency	67	169	329	197	188	103	134	84	267
Deleted by SMEs	-	-	58	-	-	-	-	195	61
AOSP Statements Used	602	508	435	369	634	506	522	354	357

Domain Consolidation									
Tasks in MOS[a]	378	166	203	304	357	338	267	188	251
Nonapplicable Systems	-	-	-	-	-	50	-	-	-
Eliminated by Doctrine	23	-	-	-	16	14	97	10	12
Collective Tasks	25	-	-	-	5	-	-	-	-
Combined Systems	57	-	-	-	-	-	-	-	-
Reserve Components Tasks	-	-	-	-	15	-	-	-	-
Tasks in Domain	273	166	203	304	321	274	170	178	239

Domain Reduction									
Tasks in Domain	273	166	203	304	321	274	170	178	239
Restricted Duty Position	44	-	42	-	-	-	-	-	-
Preliminary Sort	-	-	-	176	-	-	-	-	-
Low Frequency (High Skill Level/ AOSP Only)	53	47	-	-	90	39	-	-	-
Domain Tasks for SME Judgments	177	119	161	128	231	235	170	178	239

[a] Task list resulting from the merging of the Soldier's Manuals lists and the more detailed AOSP descriptions.

scenario identical to that used in Batch A was retained, and a "Training" scenario specifying a stateside environment and a "Combat" scenario (European non-nuclear) were developed. The 30 SMEs for each Batch B MOS were randomly divided into three groups and each group was given a different scenario as a basis for judgments.

For Batch A MOS, the judges were given the tasks on individual cards, identical to those used in task clustering, and told to rank the tasks from Most Important to Least Important. For Batch B MOS, judges were provided a list of the tasks, with descriptions, and asked to rate them on a 7-point scale from "1 = Not at all important for unit success" to "7 = Absolutely essential for unit success."

Task Performance Difficulty. To arrive at an indication of expected task difficulty, SMEs were asked to sort a "typical" group of 10 soldiers across five performance levels based on how they would expect a typical group of SL1 soldiers to perform on each task. The standard deviation of this distribution served as an index of expected performance variability.

Selection of Tasks To Be Tested

From five to nine project staff, including the individual who had prime responsibility for that particular MOS, together with six NCO/officer SMEs, participated in the task selection process for each MOS. The selection panel was provided the data summaries of the SME judgments and asked to make an initial selection of 35 tasks to represent each MOS. No strict selection rules were imposed, although the analysts were told that high importance, high performance variability, a range of difficulty, and frequently performed tasks were desirable, and that each cluster should be sampled.

The next phase was a Delphi-type negotiation among analysts to merge their respective choices into a consensus list of 35 tasks for each MOS. Information on the choices and rationale provided by each analyst in the preceding phase was distributed to all analysts, and each made a decision to retain or adjust his or her decisions, taking into account opinions others had expressed. For all MOS, three iterations were necessary.

The resulting task selection lists were mailed to each Proponent; a briefing by Project A staff was provided if requested. A Proponent representative then coordinated a review of the list by Proponent personnel designated as having the appropriate qualifications. After some minor Proponent-recommended adjustments, the final list of 30 tasks was selected for each MOS.

Assignment of Tasks to Test Mode

The initial development plan required that a job knowledge test be developed for all 30 tasks, and a hands-on test for 15 of these tasks. The considerations that constrained selection for hands-on testing were:

- Fifteen soldiers must complete all 15 hands-on tests in 4 hours.

- Scorer support would be limited to eight NCO scorers.

58

- The hands-on test site must be within walking distance of the other test activities.

- Equipment requirements must be kept within reason.

- The test must be administrable in a number of installations.

On the basis of these constraints, each of the five project analysts independently reviewed the available information and made a task selection. Following individual ratings, analysts met in group discussions and proceeded task by task to resolve differences until a consensus was reached.

Construction of Hands-On and Knowledge Tests

For both hands-on and knowledge tests, the primary source of test content was task analysis data.

Hands-On Test Development

The model for hands-on test development emphasized four activities:

- <u>Determine test conditions</u>. Test conditions were designed to maximize the standardization of the test between test sites and among soldiers at the same test site.

- <u>List performance measures</u>. Performance measures were defined as either <u>product</u> or <u>process</u> depending on what the scorer was directed to observe so as to score behavior.

- <u>State examinee instructions</u>. Examinee instructions were read verbatim to the soldier and were the only verbal communications the scorer was allowed to have with the examinee.

- <u>Develop scorer instructions</u>. These instructions told the scorer how to set up, administer, and score the test.

Job Knowledge Test Development

A multiple-choice format was selected, and 4 hours were allocated to the knowledge testing block for the field trials, to be reduced to 2 hours for Concurrent Validation testing. Allowing an average of slightly less than one minute to read and answer one item dictated an average of about nine items per task.

Knowledge test development was based on the same information that was available for hands-on development and emphasized <u>performance</u> knowledge by attempting to write items that were:

- <u>Performance-based</u>. Such items require the examinee to select an answer describing <u>how</u> something should be done. The goal was to avoid a tendency to cover information about <u>why</u> a step is done or rely on technical questions about the task or equipment. The knowledge or recall required was not to exceed what was required

59

when actually performing the task. Liberal use of quality illustrations was essential.

* **Focused on performance errors**. Performance-based knowledge tests must focus on what soldiers do when they fail to perform the task or steps in the task correctly.

Knowledge tests were constructed by project personnel with experience in test item construction and expertise in the MOS/task being tested. Test items were reviewed internally by a panel of test experts to insure consistency among individual developers.

Pilot Testing

Following construction of the tests, arrangements were made through the Proponent for troop support for a pilot test of the hands-on and knowledge tests. This procedure was conducted by the test developer. The hands-on tests involved the support of four NCO scorer/SMEs, five MOS incumbents in SL1, and the equipment dictated by the test. The knowledge tests utilized the same four NCO hands-on scorers and five MOS incumbents. The test developer went through each test, item by item, with all four NCOs simultaneously. The five incumbents took the test as actual examinees. Revisions were based on SME and incumbent inputs.

Auxiliary Instruments

Task-Specific Performance Rating Scales

Development of hands-on and knowledge tests provided two methods of measuring the sample of 15 tasks. As a third method, the soldier's peers and supervisors were asked to rate the soldier's performance on those same 15 tasks by means of a 7-point numerical rating scale. The intent was to assess performance on the same set of 15 tasks with three different methods. The rating scales were developed for administration during the field tests.

Job History Questionnaire

Although soldiers in a given MOS share a common pool of potential tasks, their actual task experience may vary substantially. To assess the likely impact of experience effects on task performance, and consequently on the Concurrent Validation strategies, a Job History Questionnaire was developed to be administered to each soldier. Specifically, soldiers were asked to indicate how **recently** and how **frequently** (in the preceding 6 months) they had performed each of the 30 tasks selected as performance criteria.

Summary

At this point the initial versions of the hands-on job sample tests and the multiple-choice knowledge tests had been developed, pilot tested, and revised. The 7-point task performance rating scales and the Job History Questionnaire had been constructed.

CRITERION DEVELOPMENT: MOS-SPECIFIC BEHAVIORALLY ANCHORED RATING SCALES

A major component of Project A criterion development was devoted to using the critical incident method to identify basic performance factors. The procedure used to identify MOS-specific performance factors was derived in large part from procedures outlined by Smith and Kendall (1963) and by Campbell, Dunnette, Arvey, and Hellervick (1973).

Development Procedure

The general development procedure involved the following steps: (a) conducting workshops to collect performance incidents for the assigned MOS, (b) editing incidents, (c) conducting the retranslation exercises, (d) developing behaviorally anchored performance rating scales (BARS), and (e) revising the scales for use in the Concurrent Validation efforts. (See Toquam et al., 1986.)

Critical Incident Workshops

Almost all participants were NCOs who were directly responsible for supervising first-term enlistees and who themselves had spent 2 to 4 years as first-termers in these MOS. Workshops for each MOS were conducted at six Continental United States (CONUS) Army posts.

Staff members first described Project A and explained the purpose of the workshop. Participants were then asked to generate accounts of Army-wide performance incidents, using examples provided as guides, and to avoid describing activities or behaviors that reflect general soldier effective-ness (e.g., following rules and regulations, military appearance), as these requirements were being identified and described in another part of the project.

After 4-5 hours, the participants were asked to identify potential job performance categories, which workshop leaders recorded on a blackboard or flipchart. Following discussion, the performance incidents written to that point were reviewed and assigned to one of the categories that appeared on the blackboard or flipchart. The remaining time was spent generating performance incidents for those categories that contained few incidents.

Results from the performance incident workshops are reported in Table 3-3 for Batch A MOS and in Table 3-4 for Batch B MOS.

Incident Retranslation and Construction of Initial Rating Scales

Evidence that the performance dimension system provides a thorough and comprehensive coverage of the critical job requirements is high agreement among judges that specific incidents represent particular components (factors) of performance, that all hypothesized factors can be represented by incidents, and that all incidents in the sample can be assigned to a factor (if they cannot, factors may be missing).

This retranslation step can also be used to develop the performance anchors for each dimension. Participants are asked to rate the level of performance described in the incident.

61

Table 3-3

BARS Performance Incident Workshops: Number of Participants and Incidents
Generated by MOS and by Location - Batch A

	MOS				Total By
Location	13B	64C	71L	95B	Location
Fort Ord					
N - Participants	14	10	5	14	43
N - Incidents	194	80	59	213	547
Mean Per Participant	13.9	8.0	11.8	15.2	12.7
..rt Polk					
N - Participants	12	15	15	15	57
N - Incidents	150	240	210	235	835
Mean Per Participant	12.5	16.0	14.0	15.7	14.7
Fort Bragg					
N - Participants	13	14	11	17	55
N - Incidents	235	221	218	225	899
Mean Per Participant	18.1	15.8	19.8	13.2	16.4
Fort Campbell					
N - Participants	13	13	10	11	47
N - Incidents	195	191	154	238	778
Mean Per Participant	11.5	13.6	17.1	15.9	14.2
Fort Hood					
N - Participants	13	13	10	11	47
N - Incidents	180	183	133	92	588
Mean Per Participant	13.9	14.1	13.3	8.4	10.7
Fort Carson					
N - Participants	19	15	13	14	61
N - Incidents	204	232	215	180	831
Mean Per Participant	10.7	15.5	16.5	12.9	13.6
Total By MOS					
N - Participants	88	81	63	86	318
N - Incidents	1159	1147	989	1183	4478
Mean Per Participant	13.2	14.2	15.7	13.8	14.1

Table 3-4

BARS Performance Incident Workshops: Number of Participants and Incidents
Generated by MOS and by Location - Batch B

	MOS					Total By Location
Location	11B	19E	31C	63B	91A	
Fort Lewis						
N - Participants	16	11	8	10	11	56
N - Incidents	211	180	124	172	130	817
Mean Per Participant	18.3	16.4	15.5	17.2	11.8	14.6
Fort Stewart						
N - Participants	14	15	15	16	16	76
N - Incidents	216	275	256	208	249	1204
Mean Per Participant	15.4	18.3	17.1	13.0	15.6	15.8
Fort Riley						
N - Participants	18	7	10	11	8	54
N - Incidents	216	123	127	133	90	689
Mean Per Participant	12.0	17.6	12.7	12.1	11.3	13.8
Fort Bragg						
N - Participants	13	14	16	15	13	71
N - Incidents	231	190	220	250	217	1,108
Mean Per Participant	17.8	13.6	13.8	16.7	16.7	15.6
Fort Sill[a]						
N - Participants	8	4	3	9	10	34
N - Incidents	26	0	13	32	20	91
Mean Per Participant	3.3		4.3	3.6	2.0	2.7
Fort Bliss[a]						
N - Participants	14	14	8	14	13	63
N - Incidents	93	70	39	71	55	328
Mean Per Participant	6.6	5.0	4.9	5.1	4.2	5.2
Total By MOS						
N - Participants	83	65	60	75	71	354
N - Incidents	993	838	779	866	761	4,237
Mean Per Participant	12.0	12.0	13.0	11.6	10.7	12.0

[a] Participants at these posts spent most of the time completing retranslation booklets rather than generating critical incidents.

63

The retranslation data were analyzed separately for each MOS. The process included computing for each incident (a) the number of raters, (b) percentage agreement among raters in assigning incidents to performance dimensions, (c) mean effectiveness rating, and (d) standard deviation of the effectiveness ratings.

The next step involved identifying those performance incidents in which raters agreed reasonably well on performance dimension assignment and effectiveness level. For each MOS, performance incidents were identified that met the following criteria: (a) at least 50 percent of the raters agreed that the incident depicted performance in a single performance dimension, and (b) the standard deviation of the mean effectiveness rating did not exceed 2.0. These incidents were then sorted into their assigned performance dimensions. Results from this sorting are presented for each MOS in Table 3-5.

Revisions After Retranslation

The categorization of the original critical incident pool produced a total of 93 initial performance dimensions for the nine MOS in Batch A and Batch B, with a range of 7-13 dimensions per MOS. Based on the retranslation results, a number of the original performance dimensions were redefined, omitted, or combined. From the original set, six were omitted and four were lost through combination. One of the omissions was due to the fact that too few critical incidents were retranslated into it by the judges. The other five were omitted because the factor represented tasks that were well beyond Skill Level 1 or were from a very specialized low-density "track" within the MOS (e.g., MOS 71L F5-Postal Clerk).

After modifying the dimension system using results from the retranslation exercise, behavioral anchors were developed for each dimension. This involved sorting effective performance incidents with mean values of 6.5 or higher, average performance with mean values of 3.5 to 6.4, and ineffective performance with mean values from 1.0 to 3.4, and then summarizing the information in each group to form three summary behavioral anchors depicting effective, average, and ineffective performance. Traditional behaviorally anchored rating scales contain specific examples of job behaviors for each effectiveness level in a performance dimension. Behavioral summary scales, on the other hand, contain anchors that represent the behavioral content of all performance incidents reliably retranslated for that particular level of effectiveness.

After the performance rating scales had been developed for each MOS, these were submitted to intensive review by the project research staff and the Scientific Advisory Group. Results from these reviews were used to clarify performance definitions and behavioral anchors.

Field Test Versions of MOS-Specific BARS

The final set of behaviorally anchored rating scales for the nine MOS for use in the field test contained from 6 to 12 performance dimensions. Each of the performance dimensions includes behavioral anchors describing ineffective, average, and effective performance. Raters were asked to use these anchors to evaluate ratees on a scale ranging from 1 (ineffective performance)

Table 3-5

Behavioral Examples Reliably Retranslated Into Each Dimension on the BARS Measures

Dimension	Number of Examples	Dimension	Number of Examples
Cannon Crewman (13B)		**Military Police (95B)**	
A. Loading out equipment	49	A. Traffic control and enforcement on post and in the field	63
B. Driving and maintaining vehicles, howitzers, and equipment	195	B. Providing escort security and physical security	128
C. Transporting/sorting/storing and preparing ammunition for fire	108	C. Making arrests, gathering information on criminal activity, and reporting on crimes	173
D. Preparing for occupation and emplacing howitzer	44	D. Patrolling and crime/accident prevention activities	236
E. Setting up communications	24	E. Promoting confidence in the military police by maintaining personal and legal standards and through community service work	118
F. Gunnery	98		
G. Loading/unloading howitzer	32		
H. Receiving and relaying communications	19		
I. Recording/record keeping	25	F. Using interpersonal communication (IPC) skills	87
J. Position improvement	14	G. Responding to medical emergencies and other emergencies of a non-criminal nature	50
	613		855
Motor Transport Operator (64C)		**Infantryman (11B)**	
A. Driving Vehicles	168		
B. Vehicle coupling	46	A. Ensuring that all supplies and equipment are field-ready and available and well-maintained in the field	73
C. Checking and maintaining vehicles	81		
D. Using maps/following paper routes	27		
E. Loading cargo and transporting personnel	75	B. Providing leadership and/or taking charge in combat situations	33
F. Parking and securing vehicles	32	C. Navigating and surviving in the field	53
G. Performing administrative duties	42		
H. Self-recovering vehicles	20	D. Using weapons safely	38
I. Safety-mindedness	80	E. Demonstrating proficiency in the use of all weapons, armaments, equipment and supplies	91
J. Performing dispatcher duties	15		
	676		
		F. Maintaining sanitary conditions, personal hygiene, and personal safety in the field	24
Administrative Specialist (71L)		G. Preparing a fighting position	29
A. Preparing, typing, and proofreading documents	183	H. Avoiding enemy detection during movement and in established defensive positions	22
B. Distributing and dispatching incoming/outgoing documents	53		
C. Maintaining office resources	73	I. Operating a radio	27
D. Posting regulations	44	J. Performing reconnaissance and patrol activities	37
E. Establishing and/or maintaining files IAW TAFFS	50		
F. Keeping records	94	K. Performing guard and security duties	76
G. Safeguarding and monitoring security of classified materials	43	L. Demonstrating courage and proficiency in engaging the enemy	
H. Providing customer service	30	M. Guarding the processing POWs and enemy casualties	15
I. Preparing special reports, documents, drafts, and other materials	19		522
J. Sorting, routing and distributing incoming/outgoing mail	28		
K. Maintaining Army Post Office equipment	2		
L. Keeping Post Office records	20		
M. Maintaining security of mail	9		
	658		

(Continued)

65

Table 3-5 (Continued)

Behavioral Examples Reliably Retranslated Into Each Dimension
on the BARS Measures

Dimension	Number of Examples	Dimension	Number of Examples
Armor Crewman (19E)		**Light-Wheel Vehicle Mechanic (63B)**	
A. Maintaining tank hull/suspension system and associated equipment	123	A. Inspecting, testing, and detecting problems with equipment	47
B. Maintaining tank turret system/ fire control system	37	B. Troubleshooting	63
		C. Performing routine maintenance	23
C. Driving/recovering tanks	80	D. Repair	101
D. Stowing and handling ammunition	39	E. Using tools and test equipment	68
E. Loading/unloading guns	30	F. Using technical documentation	56
F. Maintaining guns	43	G. Vehicle and equipment operation	18
G. Engaging targets with tank guns	45	H. Recovery	36
H. Operating and maintaining communication equipment	35	I. Planning/organizing jobs	15
		J. Administrative duties	41
I. Establishing security in the field	33	K. Safety mindedness	89
J. Navigating	11		557
K. Preparing/securing tank	27		
	504	**Medical Specialist (91A)**	
		A. Maintaining and operating Army vehicles	61
Radio Teletype Operator (31C)		B. Maintaining accountability of medical supplies and equipment	28
A. Inspecting equipment and trouble-shooting problems	50	C. Keeping medical records	31
B. Pulling preventative maintenance and servicing equipment	79	D. Attending to patients' concerns	15
		E. Providing accurate diagnoses in a clinic, hospital, or field setting	11
C. Installing and preparing equipment for operation	152	F. Arranging for transportation and/ or transporting injured personnel	44
D. Operating communications devices and providing for an accurate and timely flow of information	142	G. Dispensing medications	42
		H. Preparing and inspecting field site or clinic facilities in the field	34
E. Preparing reports	33	I. Providing routine and ongoing patient care	95
F. Maintaining security of equipment	57	J. Responding to emergency situations	142
G. Locating and providing safe transport of equipment to sites	50	K. Providing instruction to Army personnel	18
	578		511

66

to 7 (effective performance). Raters are also asked to evaluate an incumbent's overall performance across all MOS-specific performance dimensions. This final rating scale is virtually the same for all MOS; it includes three anchors depicting ineffective, average, and effective performance.

CRITERION DEVELOPMENT: ARMY-WIDE RATING SCALES

Development of Scales

The development of the Army-wide behavior rating scales (Pulakos & Borman, 1986) followed the same general procedure used for the MOS-specific BARS.

Critical Incident Workshops and Procedures

Seventy-seven officers and NCOs participated in six one-day workshops intended to elicit behavioral examples of soldier effectiveness that were not MOS-specific. A total of 1,315 behavioral examples were generated in the six workshops.

Duplicate incidents and incidents that did not meet the criteria specified (e.g., the incident described the behavior of an NCO rather than a first-term soldier) were dropped from further consideration. The remaining 1,111 examples were edited to a common format and content analyzed by project staff to form preliminary dimensions of soldier effectiveness. Specifically, three researchers independently read each example and grouped together those examples that described similar behaviors. The sorted examples were then reviewed and the groupings were revised until each author arrived at a set of dimensions that were homogeneous with respect to their content.

After discussion among project staff and with a small group of officers and NCOs at Fort Benning, a consensus was reached on a set of 13 dimensions. These were then submitted to retranslation.

Retranslation of the Behavioral Examples

The retranslation task was divided into five parts, with each part requiring a judge to evaluate 216-225 behavioral examples. Judges were provided with definitions of each of 13 dimensions to aid in the sorting, and with a 1-9 effectiveness scale to guide the effectiveness ratings.

The number of behavioral examples reliably retranslated for each of the 13 dimensions is shown in Table 3-6. The criteria established for acceptance --greater than 50 percent agreement for the sorting of an incident into a single dimension, and a standard deviation of less than 2.0 for the distribution of judges' effectiveness ratings for one incident--were met by 870 of the 1,111 examples (78%).

Two pairs of dimensions were combined because of the conceptual similarity of each of the pairs, resulting in a total of 11 Army-wide dimensions. Leading Other Soldiers and Supporting Other Unit Members were combined to form Leading/Supporting; Attending to Detail and Maintaining Own Equipment were collapsed to form Maintaining Assigned Equipment.

Table 3-6

Behavioral Examples Reliably Retranslated[a] Into Each Dimension
for Army-Wide Behavior Rating Scales

	Dimensions		Number of Examples
A.	Controlling own behavior related to personal finances, drugs/alcohol, and aggressive acts		107
B.	Adhering to regulations and SOP, and displaying respect for authority		158
C.	Displaying honesty and integrity		53
D.	Maintaining proper military appearance		34
E.	Maintaining proper physical fitness		36
F.	Maintaining living and work areas to Army unit standards		23
G.	Exhibiting technical knowledge and skill		47
H.	Showing initiative and extra effort on job/ mission/assignment		131
I.	Developing own job and soldiering skills		40
J.	Attending to detail on jobs/assignments/ equipment checks[b]	Maintaining Assigned Equipment	59
K.	Maintaining own equipment[b]		46
L.	Effectively leading and providing motivation to other soldiers[c]	Leading/ Supporting	71
M.	Supporting other unit members[c]		65
			870

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

[b]These two dimensions were subsequently combined to form a Maintaining Assigned Equipment dimension.

[c]These two dimensions were subsequently combined to form a Leading/Supporting dimension.

For each of the 11 dimensions, the reliably retranslated behavioral examples were then divided into three categories of effectiveness levels on a 9-point scale, and behavioral summary statements were written to capture the content of the specific examples at low (1-3.49), average (3.5-6.49), and high (6.5-9) performance levels.

Additional Army-Wide Scales

In addition to the 11 Army-wide BARS, two summary rating scales were prepared. First, an overall effectiveness scale was developed to obtain overall judgments of a soldier's effectiveness based on all the behavioral dimension ratings. Second, an NCO potential scale was developed to assess each soldier's likelihood of being an effective supervisor as an NCO.

Final List of Army-Wide Behavioral Rating Scales

The 11 Army-wide BARS that were retained plus the overall performance and NCO potential scales provided the following behavioral rating scales for the field test:

A. Technical Knowledge/Skill
B. Effort
C. Following Regulations and Orders
D. Integrity
E. Leadership
F. Maintaining Assigned Equipment
G. Maintaining Living/Work Areas
H. Military Appearance
I. Physical Fitness
J. Self-Development
K. Self-Control
Overall Effectiveness
NCO Potential

Development of Army-Wide Common Task Dimensions

Rating scales covering the common task domain were developed from tasks appearing in the Skill Level 1 Common Task Soldier's Manual. To develop these dimensions, a senior staff member content analyzed the specific tasks contained in the manual (e.g., Read and Report Total Radiation Dose; Repair Field Wire) and identified 13 common task areas that appeared to reflect in summary form all of the specific tasks.

Ratings consisted of evaluating how well each ratee typically performed each task on a 7-point scale. In addition, raters were given the option of choosing a "0", indicating that they had not observed a soldier performing in the task area. The 13 common task dimensions are:

A. See: Identifying Threat (armored vehicles, aircraft)
B. See: Estimating Range
C. Communicate: Send a Radio Message
D. Navigate: Using a Map
E. Navigate: Navigating in the Field
F. Shoot: Performing Operator Maintenance Weapon (e.g., M16 rifle)
G. Shoot: Engaging Target with Weapon (e.g., M16)

H. Combat Techniques: Moving Under Direct Fire
I. Combat Techniques: Clearing Fields of Fire
J. Combat Techniques: Camouflaging Self and Equipment
K. Survive: Protecting Against NBC Attack
L. Survive: Performing First Aid on Self and Other Casualties
M. Survive: Knowing and Applying the Customs and Laws of War

CRITERION DEVELOPMENT: COMBAT PERFORMANCE PREDICTION RATING SCALE

This section describes the development of a combat performance prediction scale, designed to evaluate performance under degraded conditions and the increased confusion, workload, and uncertainty of a combat environment. Two difficulties were recognized. First, although raters may often observe soldiers in garrison/field exercise performance, opportunities to observe performance under severely adverse conditions may have been limited. Second, the majority of peer and supervisor raters have never experienced combat, so they were being asked to predict how soldiers would perform in a situation that the raters themselves may not have known first-hand.

A variant of the critical incident approach was used to identify dimensions of combat effectiveness. The behavioral examples emerging from this step were content analyzed, and submitted to a retranslation and scaling procedure. Following field testing, the best items were selected and a summated rating scale format was developed, which was used in the Concurrent Validation.

Critical Incident Workshops. Forty-six officers and NCOs participated in one of four one-day critical incident workshops. All participants were combat veterans, the large majority with experience in Vietnam. In each workshop, a staff member first described Project A and explained how the prediction of combat performance was an integral part of the project. The workshop leader next presented a preliminary set of literature-based dimensions of combat effectiveness, and possible modifications and additions were discussed. The rest of each workshop was devoted to writing and reviewing the examples.

A total of 361 examples of positions and negative behavior was generated in the four workshops. After duplicates and items that were specific to officers, MOS, or equipment were eliminated, 158 usable examples remained. A review of the critical incidents that had been used in the Army-wide rating scale retranslation workshops revealed 73 that described behavior in a combat-type situation, such as behavior under adverse conditions during training and field exercises. These examples were added to the 158 usable examples from the combat workshops. The distribution is shown in Table 3-7.

Three staff members independently read each example and grouped those that described similar behaviors. The content analysis of the incidents resulted in a reduction of the number of dimensions from 11 to 8. The revised dimensions are shown in Figure 3-1. Employing the eight dimensions and 231 behavioral examples, materials were developed for retranslation and scaling workshops.

70

Table 3-7

Number of Edited Examples of Combat Behavior

Type of Behavior	Combat Workshops	Army-Wide Workshops	Total
Positive	96	42	138
Negative	62	31	93
Total	158	73	231

Retranslation and Scaling Workshops. In the retranslation process, acceptable agreement was defined as greater than 50 percent of the 16 judges sorting an example into the same dimension. Of the 231 examples, 108 did not meet this criterion. The workshop participants also rated each incident in terms of how well it would discriminate the "best" from the "worst" performer under adverse conditions. For the summated scale form of the Combat Performance Prediction Scale, the goal was to select items that represented the domain of combat effectiveness and discriminated between performance extremes. The summated scale form was used to anchor performance rating scales with more general or abstract behavioral examples. These general statements of performance at different levels of effectiveness add perspective to the depiction of performance (Borman, 1986, p. 105).

Allowing for time constraints in testing, and eliminating poor items, 80 items were selected. To reduce the administrative burden on any one rater, two forms (Form A and Form B) were developed. Each contained 60 items--40 common to both forms and 20 unique to one form.

Review and Rescaling. The two proposed 60-item forms of the Combat Performance Prediction Scale were reviewed by three company grade Army officers and three ARI scientists. As a result of that review, three items common to both forms were deleted and a large proportion of the remaining 77 items were reworded. Since the rewording was extensive, the 77 items were subjected to a rescaling, using the same workshop procedures as for the original scaling. Eight officers and one civilian (seven of the nine were combat veterans) made the "best" and "worst" combat soldier ratings for each of the 77 items. Only one item was dropped, because it did not discriminate between effective and ineffective.

CRITERION DEVELOPMENT: ADMINISTRATIVE/ARCHIVAL RECORDS

A major activity within the overall program of performance criterion development was to explore the use of the archival administration records as first-tour job performance criteria and in-service predictors of soldier effectiveness (Riegelhaupt, Harris, & Sadacca, 1985). The Enlisted Master File (EMF), the Official Military Personnel File (OMPF), and the Military Personnel Records Jacket (MPRJ) are the Army records sources that contain administrative actions that could be used to form measures of first-tour soldier effectiveness.

71

A. Cohesion/Commitment to Others
   ● Ability and desire to foster a common spirit of devotion and
     enthusiasm among members of a group
   ● Concern for the physical/emotional welfare of the individual
     members of the group
   ● Commitment to maintaining/enhancing the effectiveness of the group

B. Intelligence/Common Sense
   ● Ability to learn quickly and apply the newly acquired knowledge/
     skill in a novel situation
   ● Ability to size up a situation and use available resources to make
     a decision
   ● The exercise of appropriate judgment

C. Self-Discipline/Responsibility
   ● Willingness to accept responsibility for the accomplishment of the
     task at hand
   ● Concern for conditions that jeopardize the safety of self and
     others
   ● Concern for the maintenance of weapons and equipment, etc.

D. Physical/Medical Condition
   ● Ability and willingness to maintain both physical and medical
     fitness
   ● Physical endurance as demonstrated by little or no reduction in
     performance even after or during prolonged or strenuous activities
   ● Concern for proper health care/hygiene to avoid sickness and
     disease

E. Mission Orientation
   ● Willingness to make sacrifices and endure hardships to accomplish
     mission
   ● Commitment and dedication to accomplishing one's assigned duties/
     responsibilities
   ● Willingness to accept a reasonable amount of risk in the pursuit
     of mission accomplishment

F. Technical/Tactical Knowledge
   ● Ability to follow SOP
   ● Knowledge of and ability to coordinate weapons, ammunition, and
     equipment
   ● Ability to perform MOS-specific and common soldiering tasks

G. Psychological Effects of Combat
   ● Reaction to stress associated with shooting and killing, losing a
     unit/team leader, seeing others wounded or killed, waiting for
     orders between engagements, etc.
   ● Ability to perform duties with little or no decrement under
     emotionally stressful situations

H. Initiative
   ● Ability and willingness to take the appropriate action at the
     appropriate time without being told to do so

Figure 3-1. Revised set of combat performance dimensions.

The EMF is an automated inventory of personal data, enlistment conditions, and military experience for every enlisted individual currently on the U.S. Army payroll. It contains a large number of variables for each individual, ranging from pay grade to Skills Qualification Test (SQT) scores to the Army's operational performance appraisal ratings in the form of the Enlisted Efficiency Report (EER).

The OMPF is the permanent, historical, and official record of a member's military service. The information for enlisted personnel is maintained on microfiche records at the Enlisted Records and Evaluation Center, Fort Benjamin Harrison.

The MPRJ, or 201 File, is the primary mechanism for storing information about an individual's service record. Updates/additions/corrections to the file are made at the time of the action. The MPRJ physically follows the individual wherever he or she goes and is normally located at the Military Personnel Office (MILPO) that serves the soldier's unit.

A series of small pilot tests were conducted to explore the information content of each source, identify the problems that would be involved in using it, and develop an appropriate data collection protocol that could be used in a large-scale systematic records search. In so doing, an initial list of potentially useful administrative records was identified, and is shown in Table 3-8.

Comparative Pilot Test

A systematic comparison of the three data sources was carried out on a pilot sample of 650 records. The original plan was to collect data from the MPRJ for a sample of 750 soldiers, 150 in each of five MOS at five Army posts. To achieve this sample size, the records of 200 soldiers at each post were requested. Data were collected by teams of two research staff members in 2-day visits to each of five posts. Only those soldiers who entered the Army between 1 July 1981 - 31 July 1982 at an initial grade of PFC or less were retained. The result was a sample of 650 soldiers in the 05C, 11B, 64C, 71L, or 91B MOS who had been in the Army between 14 and 27 months.

Military Personnel Records Jacket (MPRJ) - Official Military Personnel File (OMPF) Comparison

Using the records collection form developed to extract records data from the MPRJ, three research staff members spent 2 days collecting records data from the OMPFs of 292 soldiers. The 292 individuals represented a random sample of the 650 soldiers from whose MPRJs administrative records data had previously been collected. The MPRJ was found to be a much richer source than the OMPF for information on the administrative actions of interest in Project A. In the extreme case, even information relevant to a soldier's reenlistment eligibility was not available from the OMPF.

73

Table 3-8

List of Administrative Measures Indicative of Soldier Effectiveness

---

- Comparison of Skill Level of Primary to Duty MOS.
- Existence of Secondary MOS.
- Existence of Skills Qualification Identifier (SQI).
- Existence of Additional Skill Area (ASI).
- Existence of Language Identifier.
- Record of Skill Qualification Test (SQT) Score Within Past 12 Months.
- Type of Reenlistment Eligibility.
- Type of Military Education Leadership Course.
- Level of Highest Civilian Education.
- Promotion Rate.
- Existence of Promotion Packet at E4.
- Number and Type of Awards/Badges.
- Record of Requalification Weapons Score Within Past 12 Months.
- Number and Type of Certificates of Achievement/Appreciation/ Commendation.
- Number and Type of Letters of Appreciation/Commendation.
- Number and Type of Letters of Reprimand/Admonition.
- Number of Additional Civilian Education Classes Completed.
- Number and Type of Correspondence Courses Completed.
- Number of Additional Civilian Education Classes Completed.
- Course Summary and Abilities Ratings - Service School.
- Professional Competence and Standards Ratings and Summary Score of Enlisted Efficiency Report.
- Type, Sentence, Suspension, Vacation of Courts-Martial.
- Existence of Courts-Martial Proceedings in Action Pending.
- Reason for Bar to Reenlistment.
- Number and Duration of AWOL.
- Number of Violations, and Reason for Articles 15.
- Reason for FLAG Action.
- Number of and Reason for Disposition - Block to Promotion.

Military Personnel Records Jacket (MPRJ) - Enlisted Master File (EMF) Comparison

Unlike the MPRJ-OMPF comparison, a rather high degree of correspondence existed between the MPRJ and the EMF. Even in light of delays in data entry, the correspondence between sources was impressive and highlighted the benefits of having current EMF information available.

Variable Selection

A first step in determining the usefulness, for Project A purposes, of the administrative variables collected from MPRJs (201 Files) was to select those measures with an acceptable amount of variance. Based upon the frequency distributions and intercorrelations of the possible indexes, and regulations governing reenlistment and promotion criteria, six variables were

74

selected as having the highest potential for being useful criteria and in-service predictors for Project A:

- Eligible to Reenlist
- Number of Letters/Certificates
- Number of Awards
- Number of Military Training Courses
- Has Received Article 15/FLAG Action
- Promotion Rate (Grades Advanced/Year)

Relationships of Administrative Measures With Other Variables

Each of the six administrative measures and a combined "Has Received Letter/Certification/Award" variable were subjected to a series of analyses. These included an examination of MOS and Post differences; stepwise multiple regressions, in which AFQT, Moral Waiver, Sex, and Race were entered after controlling for Post and MOS effects; and univariate analyses, in the form of chi-square tests, for those variables entered into the regression equation with a significant F value at the time of first entry.

First, there was no evidence that a soldier's race was a significant determiner of his or her Reenlistment Eligibility, Number of Awards, or any other of the Army-wide administrative measures. Second, although a soldier's sex was related to Awards (males received more) and to Letter/Certificate (females received more), when the two variables were combined into the Letter/Certificate/Award measure, sex differentials were no longer statistically significant.

Third, Armed Forces Qualification Test (AFQT) score or mental category (see Appendix A) was related to successfully completing Military Training Courses and to Number of Awards, indicating the possible usefulness of the ASVAB in predicting aspects of Army-wide performance. Fourth, both Reenlistment Eligibility and Promotion Rate (from E-1 to E-4), which may be related to non-cognitive as well as cognitive factors, do not appear to be dependent on the soldier's location (Post), MOS, or demographic group (i.e., these measures seem to be fairly even-handedly administered Army-wide).

Finally, there were distinct MOS and post differences in average scores for most of the measures. For example, Administrative Specialists (71L) received more letters/certificates and Infantrymen (11B) more awards than soldiers in other MOS. Soldiers at one of the five posts visited received more letters, certificates, and awards, and more extra training than soldiers at the other posts. Care should be exercised in pooling performance measurement data across MOS and posts.

Criterion Field Test: Self-Reports of Administrative Actions

While the use of administrative measures is consonant with the Project A multimethod approach to performance measurement, and while these indexes hold predictors of second-tour performance, it must be asked whether the effort and expense of collecting these indexes from the 201 Files are justified by the outcome. Also, while there was a high degree of correspondence between information on the EMF computerized file and information collected from the individual 201 Files, a number of the most promising variables were not available from the EMF.

75

Accordingly, another method of obtaining information was tried out. A self-report instrument, the Personnel File Information Form, was developed and administered during the Batch A field testing. The self-report information could then be compared to the information in actual 201 Files, obtained by the project team during the field test period.

CRITERION DEVELOPMENT: MEASURES OF TRAINING SUCCESS

Training achievement tests were developed to measure training success for the 19 MOS in the Project A sample (Davis, R. H., Davis, G. A., Joyner, & deVera, 1986). The training performance measures were to serve both as criteria for selection/classification predictor validation and as in-service predictors of later job performance. A longstanding question is whether training performance criteria and job performance criteria provide the same information about predictor validity.

Within the Army, there is a very close relationship between training content and tasks performed on the job. As a matter of doctrine, training must be job-related, and the knowledges and skills necessary for the performance of a job at Skill Level 1 are taught in Advanced Individual Training. As a result, if content validity is based on curricular materials alone, then by design most of the items should be job-related.

There are perhaps three critical components of content validity in this context. First, the content domain should be clearly defined and the boundaries of the domain from which test content is drawn should be clearly understood. Once the boundaries are defined, experts should be able to agree as to whether or not items fall inside or outside of those boundaries. For training content, the domain was described by the Program of Instruction (POI) lesson plan, technical publications, and training manuals. For the job, content was specified by Army Occupational Surveys, technical publications, Soldier's Manuals, and the Common Task Manual. Second, the sample of content to be tested should be representative of the domain. Third, the content to be tested should be highly relevant for the goals of training.

Also, it seems clear that some trainees learn relevant knowledges and skills that are not part of the explicit goals of instruction and go beyond the formal course content. From the perspective of criterion development, the most successful trainee is one who goes beyond the formal course objective. This is a distinction between _direct_ and _incidental_ learning. A relevant question is the degree to which the correlation between training performance and job performance is a function of direct learning during training, incidental learning during training, or individual differences in basic abilities that are present before training starts.

Test Development Procedure

The principal steps in the construction of the training achievement tests were as follows:

- Preparation of the item "budget" to ensure coverage of duty areas per MOS
- Development of the initial item pool
- Review of item pool by job incumbents

- Review of item pool by school trainers
- Pilot administration of items to trainees
- Preparation of the item pools for administration to job incumbents
- Administration to job incumbents (Field Tests)
- Review by TRADOC Proponent agencies
- Preparation of the item pools for administration to job incumbents in the Concurrent Validation

Although each test went through many revisions during this process, there were three principal versions: (a) the initial item pool, (b) the version administered to incumbents in the field test, and (c) the version administered to incumbents in the Concurrent Validation. Figure 3-2 summarizes the developmental procedures and illustrates the difference in procedures for Batch A/B and Batch Z.

Development of the Initial Item Pool

The initial content source was the Army Occupational Survey Program (AOSP) which uses a questionnaire checklist of several hundred items to survey job incumbents about specific job tasks that they do or do not perform. Related tasks are combined into duty areas and the number of duty areas in each of the 19 MOS ranged from 15 to 23. A key statistic reported is the percentage of soldiers at different skill levels who are performing the task activity.

Before the AOSP items were used, 99 percent confidence intervals were computed for the mean percentage performing each task, and tasks equal to or less than the lower boundary of the confidence interval were deleted. The remaining task statements were then reviewed by 4-6 SMEs for relevance and clarity and, using the following procedure, an item budget was drafted with an initial target of 225 items.

The match between AOSP duty areas and training objectives was determined by preparing a matrix of the AOSP duty areas versus the subdivisions of the POI. Three outcomes were possible: (a) duty areas matched Army training lessons completely; (b) duty areas did not match any training lesson; (c) training lessons did not match any duty area. The majority of the item budget, 200 items, was allocated to the first two categories.

Items were then budgeted in proportion to how much they were emphasized in training: The greater the overlap between the AOSP tasks (within a duty area) and the training objectives (within the POI), the more items were written to represent job/training content. The remaining items (out of the original 200) were assigned to job-only content.

After item budgets were established, written materials dealing with job training activities were examined and multiple-choice items were drafted for all MOS. The item-writing group included the research staff and contract item-writers.

77
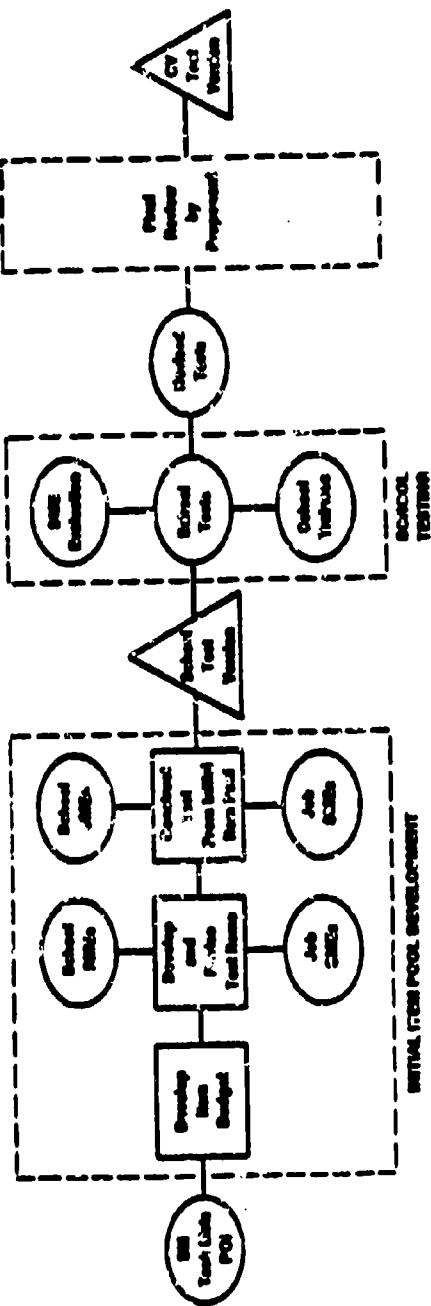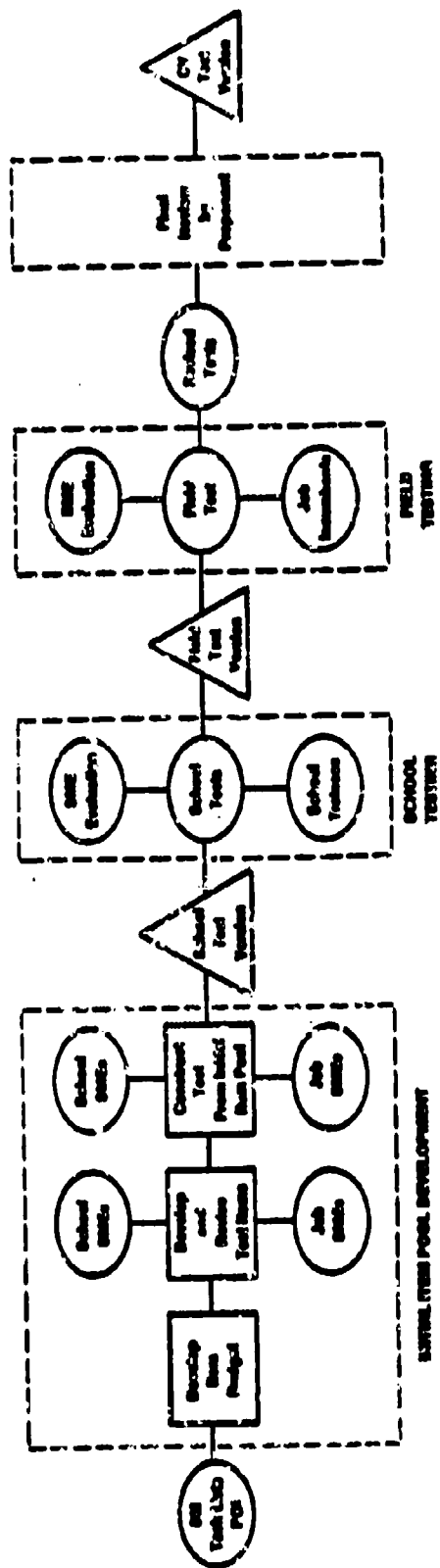
Figure 3-2. Development process for tests of achievement in training.

Review by Job Incumbents

After the pool was first reviewed by one subject matter expert who purged the item pool of its more glaring faults, the items were then reviewed by job incumbents for accuracy and relevance during a series of site visits, and items were revised where appropriate. Incumbents were next asked to rate the importance of each item on a 5-point scale in three different contexts: combat, combat readiness, and garrison duty.

Mean interrater reliabilities were reasonably high for the combat and combat readiness scenarios, .74 and .71 respectively, but somewhat lower for the garrison scenario, .60. To establish the relevance of the draft test items, incumbents were asked, "Do Skill Level 1 personnel in this MOS need to use this knowledge on the job?"

Review by School Trainers

The item pool was also reviewed by trainers at one of the training sites for the MOS. As with the review by job incumbents, the trainers reviewed items for technical accuracy and appropriate vocabulary, and rated item content for importance and relevance to the goals of training. It was during such site visits that pilot tests were conducted with trainees, as described in the next subsection.

Administration to Trainees

After review by job incumbents and trainers, test items were administered to groups of trainees in their last week of training. A sample of trainees was also interviewed after the test to obtain information about the clarity and comprehensibility of the items.

Preparation of Batch A and Batch B Training Achievement Tests for Field Tests With Job Incumbents

After all the SME judgments were made and trainee tryouts completed, the items were revised in accordance with the SME and trainee comments and the item pools were prepared for administration to job incumbents in the field tests. Data from the field test administration were later used to convert the pools of draft items into the standardized training knowledge tests.

As the item pools were cut and items added or changed in these early test construction steps, items were dropped if they were judged to be of little importance or no relevance. However, the nature of the item budget was preserved by adding new items if necessary.

Field Test Instruments

At this stage the nine training achievement tests for the MOS in Batch A and Batch B were deemed ready for field testing with job incumbents.

Up to this point the 10 tests for the 10 MOS in Batch Z followed the same developmental steps as for the tests in Batches A and B. However, as noted previously, the Batch Z instruments were not field tested with job incumbents. Consequently, the Concurrent Validation versions of the 10 tests retained more items than do the nine A/B tests. Additional item analyses were

79

carried out for Batch Z on the basis of the data from the Concurrent Validation sample.

CRITERION FIELD TESTS

The complete array of specific criterion measures evaluated in the criterion field test is given below. Again, the distinction between MOS-specific and Army-wide is that the latter are the same across all MOS. The content of the MOS-specific measures, regardless of whether they are job samples, knowledge tests, or ratings, concerns a particular job and is based on the task content of that job. Also, the judgment (i.e., rating) of "NCO potential" refers to a first-tour enlisted soldier's potential, assuming the individual would reenlist, for being an effective noncommissioned officer, with supervisory responsibilities, during the second tour of duty.

MOS-Specific Performance Measures

1) Paper-and-pencil tests of achievement during training, consisting of job-relevant knowledge tests of 100 to 200 items per MOS. Items can be aggregated by POI module or by MOS duty area.

2) Paper-and-pencil tests of knowledge of task procedures consisting of an average of about nine items for each of 30 major tasks for each MOS. Item scores can be aggregated in at least four ways.

   - Sum of item scores for each of the 30 tasks.
   - Total score for 15 tasks also measured hands-on.
   - Total score for 15 tasks not measured hands-on.
   - Total score on all 30 tasks.

3) Hands-on measures of proficiency on tasks for each MOS, measured on 15 tasks selected from the 30 tasks measured with the paper-and-pencil test.

   - Individual task scores.
   - Total score for all 15 tasks.

4) Ratings of performance, using a 7-point scale, on each of the 15 tasks measured via hands-on methods by:

   - Supervisors
   - Peers
   - Self

5) Behaviorally anchored rating scales of 6-12 performance dimensions for each MOS by:

   - Supervisors
   - Peers
   - Self

6) A general rating of overall MOS task performance by:

- Supervisors
- Peers
- Self

7) A job history questionnaire administered to incumbents to determine the frequency and recency of task performance on the 30 tasks for which job knowledge tests were developed.

Army-Wide Measures

1) Eleven behaviorally anchored rating scales designed to assess the dimensions listed below. Three sets of ratings (i.e., from supervisors, peers, and self) were obtained on each scale for each individual.

- Technical Knowledge/Skill
- Initiative/Effort
- Following Regulations/Orders
- Integrity
- Leading and Supporting
- Maintaining Assigned Equipment
- Maintaining Living/Work Areas
- Military Appearance
- Physical Fitness
- Self-Development
- Self-Control

2) A rating of general overall effectiveness as a soldier by:

- Supervisors
- Peers
- Self

3) A rating of noncommissioned officer potential by:

- Supervisors
- Peers
- Self

4) A rating of performance on each of 14 common tasks from the Manual of Common Tasks by:

- Supervisors
- Peers
- Self

5) A 77-item summated rating scale of expected combat effectiveness.

- Supervisors
- Peers
- Self

81

6) A 14-item self-report measure (the Personnel File
Information Form) of certain administrative indexes
such as awards, letters of recommendation, and
reenlistment eligibility.

7) The same administrative indexes taken from 201
Files (by project staff).

8) An Environmental Questionnaire, a descriptive questionnaire
completed by both incumbents and supervisors for the
purpose of describing 14 factors pertaining to organizational
climate, structure, and practice (Peterson, Hough, Ashworth,
& Torquam, 1986).

9) A 99-item Leader Behavior Questionnaire to measure incumbents'
perceptions of the leadership behaviors and practices
in their unit (White, Gast, & Rumsey, 1986).

10) A Measurement Method Questionnaire administered at the
end of the testing sessions to obtain soldiers' reactions to
the various types of testing.

Samples

The samples for the field tests were drawn from the nine Batch A and
Batch B MOS and from six different locations. Tables 3-9 and 3-10 provide a
breakdown of the criterion field test sample sizes by MOS and location, and by
race and sex, respectively. The USAREUR data collection site was just outside
Frankfurt, Germany.

Table 3-9

Field Test Sample Soldiers by MOS and Location

Location	MOS									Total
	11B	13B	19E	31C	63B	64C	71L	91A	95B	
Fort Hood	--	--	--	--	--	--	48	--	42	90
Fort Lewis	29	--	30	16	13	--	--	24	--	112
Fort Polk	30	--	31	26	26	--	60	30	42	245
Fort Riley	30	--	24	26	29	--	21	34	30	194
Fort Stewart	31	--	30	23	27	--	--	21	--	132
USAREUR	58	150	57	57	51	155	--	58	--	596
Total	178	150	172	148	156	155	129	167	114	1,369

Table 3-10

Field Test Sample Soldiers by Gender and Race

Race	Male	Female	Total
Black	330	58	388
Hispanic	37	3	40
White	789	104	893
Other	43	5	48
Total	1,199	170	1,369

Procedure

For the purpose of data collection in the field tests, the criterion measures were divided into four major blocks corresponding to:

(1) Hands-on (job sample) measures (HO).
(2) Rating measures (R) - both Army-wide and MOS-specific.
(3) Paper-and-pencil measures of job knowledge (K_s).
(4) Paper-and-pencil measures of training achievement (K_j).

Each block comprised one-half day of participant time and each participant was tested for a 2-day period.

During the week preceding data collection at each research site, the scorers for the hands-on (job sample) measure were given 2 days of training on scoring procedures, test standardization, and the overall design and objectives of Project A.

Analysis

The general data analytic steps were straightforward and consisted of the following:

(1) An item analysis summary table for each knowledge test for each MOS. The table for each MOS summarized item discrimination indexes, item difficulties, and the frequency of items that were flagged for various kinds of potential keying errors (o.g., negative correlation with total score, high frequency of response for incorrect answer).

(2) An item (where task = item) analysis for each hands-on (job sample) test.

(3) Frequency distribution and scale statistics for each rating scale for each MOS.

(4) Interrater reliabilities for the individual rating scales.

(5) Split-half correlations (Spearman-Brown estimates) for the knowledge tests and hands-on measures, test-retest coefficients for the hands-on measures, and internal consistency indexes where applicable.

(6) A complete intercorrelation matrix of all the criterion variables for each MOS down to the scale score and task score level (i.e., the matrix included all the variables listed in the previous sections).

(7) A set of reduced intercorrelations matrixes that included subsets of the total array of variables.

(8) Factor analyses for selected matrixes, primarily those having to do with the rating scale measures.

The results of the above analyses were prepared in a master data book for each MOS. Each data book contained item and scale analyses, inter-correlations down to the scale and subscale level, and factor analyses of selected data sets.

These data were then carefully scrutinized by a designated criterion analysis group. The group included the principal investigator for each of the criterion measures, the principal scientist for the project, the ARI chief scientist and task monitors for the project, and the assistant project director, who served as chair.

The objectives of the group were to review the results of the field tests and agree upon the specific revisions to be made in each criterion measure before the criterion array was declared the set of criterion measures that would be used for the Concurrent Validation.

FIELD TEST RESULTS

Job Knowledge Tests

Between 14 and 18 percent of the items in each MOS item set were revised as a consequence of field test experience, and between 17 and 24 percent of the items were dropped. The median difficulty levels were 55 to 58 percent for five of the MOS, with the MOS 63B, 91A, 19E, and 95B tests having medians of 65 to 74 percent. Although some skew in item difficulties was observed, it was not extreme.

The means, standard deviations, and reliabilities for the total test score in each MOS are shown in Table 3-11. The reliabilities are split-half coefficients, using 15 task tests in each half, corrected to a total length of 30 task tests.

Table 3-11

Means, Standard Deviations, and Split-Half Reliabilities for
Knowledge Test Components for Nine MOS

MOS	Mean (%)	Standard Deviation	Split-Half Reliability[*]
13B - Cannon Crewman	58.9	12.6	.86
64C - Motor Transport Operator	60.3	10.1	.79
71L - Administrative Specialist .	55.8	10.4	.81
95B - Military Police	66.4	9.2	.75
11B - Infantryman	56.0	10.5	.91
19E - Armor Crewman	64.0	10.1	.90
31C - Single Channel Radio Operator	57.7	9.6	.84
63B - Light Wheel Vehicle Mechanic	64.4	9.1	.86
91A - Medical Specialist	69.8	8.1	.85

[*]Fifteen task tests in each half, corrected to a total length of 30 tests.

Hands-On Tests

The hands-on tests resulted in 15 task scores, with each task composed of a number of scorable steps. Steps that had low or negative correlations with the total task score were reviewed to identify situations where performance prescribed by local practices was as correct at that site as doctrinally prescribed procedures. Instructions to scorers and to soldiers were revised as necessary to insure consistent scoring.

However, use of step statistics to revise task tests was purposely limited because a task test usually represents an integrated procedure and removal of a step which the Soldier's Manual specifies as a part of the job may result in deleting a doctrinal requirement. Table 3-12 shows, for each MOS, the means, standard deviations. and split-half reliability estimates of the hands-on components across revised task tests.

In revising the hands-on tests, the goal for each MOS was a set of between 14 and 17 task tests. Field test experience indicated that reductions of this magnitude would meet the time allotments for Concurrent Validation. Both the field test results and additional systematic judgments by the project staff of the "suitability" of the test for hands-on measurement were used to make these reductions.

85

Table 3-12

Means, Standard Deviations, and Split-Half Reliabilities for
Hands-On Test Components for Nine MOS

MOS	N	Mean (%)	Standard Deviation	Split-Half Reliability[a]
13B - Cannon Crewman	146	54.5	14.0	.82
64C - Motor Transport Operator	149	72.9	9.1	.59
71L - Administrative Specialist	126	62.1	9.9	.66
95B - Military Police	113	70.8	5.8	.30
11B - Infantryman	162	56.1	12.3	.49
19E - Armor Crewman	106	81.1	11.8	.56
31C - Single Channel Radio Operator	140	80.1	10.7	.44
63B - Light Wheel Vehicle Mechanic	126	79.8	8.7	.49
91A - Medical Specialist	159	83.4	11.4	.35

[a]Calculated as 8-test score correlated with 7-test score, corrected to 15
tests.

The extent of the changes made on the tests, considering both obtained data
and informed judgments, was small. Among common task tests, judgments of hands-
on suitability resulted in deleting five tests. Additionally, in each MOS two to
five MOS-specific tasks were dropped.

Proponent Agency Review

Following on the adjustment steps described above, each MOS was covered
by a set of 15-17 hands-on tests, and a set of knowledge items that was 60 to
70 percent of the set that had been field tested. The array of hands-on and
knowledge tests for each MOS is summarized in Table 3-13.

The final step in the development of hands-on and knowledge tests was
Proponent agency review. This step was consistent with the procedure of
obtaining input from Army subject matter experts at each major developmental
stage.

The Proponent was asked to consider two questions: (a) Do the measures
reflect doctrine accurately, and (b) do the measures cover the major aspects
of the job? A Proponent representative was given copies of the measures;
staffing of the review was left to the discretion of the Proponent agent.

Table 3-13

Summary of MOS Task Tests Before Proponent Review

MOS	Total Hands-On	Knowledge Items
13B	17	177/181
64C	16	168
71L	15	148
95B	15	210
11B	15	198
19E	15	196
31C	15	215
63B	15	196
91A	15	234

Item changes by Proponents generally affected fewer than 10 percent of the items within an MOS and most such changes involved the wording, not the basic content, of the item. Changes affecting the task list occurred in only three MOS.

In determining whether any of these task list changes constituted a major shift in content coverage, special consideration was given to the principle, applied in the initial task selection, that every cluster of tasks be represented by at least one task. For MOS 71L and MOS 95B, each cluster was still represented after the Proponent changes had been implemented. For MOS 11B, the deletion of Perform PMCS on Tracked or Wheeled Vehicle and Drive Tracked or Wheeled Vehicle left one cluster, consisting of tasks associated with vehicle operation and maintenance, unrepresented. However, the Infantry School's position was that tasks in this cluster did not represent the future orientation of the 11B MOS, so this omission was considered acceptable under the selection criteria.

A second condition in which strict adherence to Proponent suggestions was not necessarily advisable was where the suggestions could not be easily reconciled with documented Army doctrine. Where conflict with documentation emerged, the discrepancy was pointed out; if the conflict was not resolved, items were deleted.

87

Finally, if Proponent comments seemed to indicate a misunderstanding of the intended purpose or content of test items, clarification was attempted. The basic approach was to continue discussions until some mutually agreeable solution could be found.

Task Performance Rating Scales

Inspection of the task performance rating data revealed large level differences in the mean ratings provided by two or more raters of the same soldier, and reliabilities varied widely across the tasks. During the Batch A field tests, it was observed that supervisors and peers, confronted with only the task title, might not have been entirely clear on the scope of tasks they were rating. Low interrater reliability supported this observation. Consequently, for the Batch B data collection for two MOS (31C and 19E), the task statements were augmented with the brief descriptions of the tasks that had been developed for the task clustering phase of development. However, this modification did not appear to affect results from these MOS.

MOS-Specific Ratings (BARS)

For each MOS, the reliability estimates computed for performance dimension ratings provided by supervisors were compared with estimates for dimension ratings provided by peers to identify problem dimensions. (See Table 3-14 for a summary of the median reliability estimates as well as the range of reliabilities for each MOS.)

Revisions Based on Field Test Data

For most MOS, there appears to be no consistent pattern when reliability estimates computed for supervisor ratings are compared with those computed for peer ratings. Within MOS 95B one performance dimension, Providing Security, appeared to present problems for both rater groups. The interrater reliability estimate computed separately for supervisors and peers was .39. Therefore, the definition as well as the behavioral anchors for this particular dimension were clarified.

For the remaining MOS-specific rating scales, performance dimensions with low reliability estimates for supervisor or peer ratings were identified. The rating scale definitions and anchors developed for these dimensions were reviewed, and revised if it seemed appropriate. Since very little leniency or central tendency error was exhibited, no changes were made in the scales as the result of these data.

Revisions Based on Proponent Review

For one MOS, Military Police (95B), the Proponent asked for more extensive changes. Incumbents in this MOS provide combat and combat support functions. Therefore, four performance dimensions describing these requirements were added to the MOS-specific rating scales: Navigation (Dimension H); Avoiding Enemy Detection (Dimension I); Use of Weapons and Other Equipment (Dimension J); and Courage and Proficiency in Battle (Dimension K). Definitions and behavioral anchors for these scales had been developed for the Infantryman (11B) rating scales. Proponent representatives reviewed these definitions and anchors and authorized including the same information in the Military Police performance rating scales.

Table 3-14

Summary of Reliability Estimates of MOS-Specific BARS for Supervisor and Peer Ratings

MOS		Supervisors			Peers	
		Median	Range		Median	Range
13B	Cannon Crewman	.54	.33 .76	J. Position Improvement K. Overall Performance	.54	.48 .64	B. Receiving and Relaying Communications C. Load/Unload Howitzers
64C	Motor Transport Operator	.57	.47 .64	F. Parking and Securing Vehicles I. Safety Mindedness	.54	.32 .64	C. Performing Administrative Duties D. Using Maps and following proper Routes
71L	Administrative Specialist	Not calculated - Insufficient Number of Supervisor Raters per Ratee			.46	.37 .55	B. Providing Customer Service C. Safeguarding and Monitoring Security of Classification Documents I. Overall Performance
95B	Military Police	.55	.39 .74	D. Providing Security H. Overall Performance	.65	.39 .71	B. Providing Security H. Overall Performance
11B	Infantryman	.53	.29 .63	L. Prisoners of War A. Maintaining Supplies, Equipment and Weapons	.55	.38 .44	C. Avoiding Enemy Detection C. Navigation
19E	Armor Crewman	.57	.44 .73	E. Maintaining Guns F. Engaging Targets with Tank Guns	.43	.29 .63	C. Storing Ammunition Aboard Tanks I. Overall Performance
31C	Radio Teletype Operator	.63	.57 .70	C. Operating Communication Devices G. Overall Performance	.60	.52 .65	A. Inspecting and Servicing Equipment G. Overall Performance
63B	Light-Wheel Vehicle Mechanic	.62	.43 .67	C. Performing Routine Maintenance L. Overall Performance	.59	.35 .70	E. Recovery C. Performing Routine Maintenance
91A	Medical Specialist	.66	.45 .75	C. Providing Routine and Ongoing Patient Care C. Keeping Medical Records	.62	.44 .68	F. Preparing and Inspecting Field Site or Clinic Facilities I. Providing Health Care Instruction to Army Personnel

89

Army-Wide Rating Measures

Analyses of the field test data from the Army-wide rating measures focused on (a) distributions of the ratings, (b) interrater reliabilities, and (c) intercorrelations among the rating scale dimensions.

Findings suggest that raters did not exhibit excessive leniency or central tendency. Means were generally between 4 and 5 on the 7-point scale. Reliabilities of the individual behavioral scales were respectable (.51 - .68, median = .58) and composites of individual scales would be higher. The single-scale Overall Effectiveness and NCO Potential reliabilities were likewise reasonably high (median = .66). Regarding the Army-wide common task ratings, interrater reliabilities for the common task scale interrater reliabilities were lower (.33 - .60, median = .44). Supervisor and peer ratings had very similar levels of interrater reliability.

Overall, the rating scale intercorrelations were not as high as are usually found and were substantially lower than the individual scale reliabilities. This is particularly significant because the scale reliabilities (i.e., the intraclass r) incorporated rater differences as error while the scale intercorrelations did not (i.e., all correlations were based on the same set of raters).

As with the MOS-specific BARS scales, experience administering the Army-wide rating scales during Batch A indicated that some soldiers had difficulty with the amount of reading required. In addition, a few of the statements anchoring the different effectiveness levels appeared to be multidimensional.

Between the Batch A and Batch B administrations, one of the 13 common task scales was dropped because a 13th scale would have required an additional page on the printed version of the scales. The task dimension that had the lowest interrater reliability and seemed the most redundant with others was eliminated for Batch B and the Concurrent Validation.

Finally, after the instruments were submitted to Proponent review, the Army-wide effectiveness dimension Maintaining Living/Work Areas was dropped to reduce the time required to complete these scales. Experts judged that dimension to be the least important and the most expendable.

In summary, only minimal changes were made to the Army-wide rating scales as a result of the field tests: first, eliminating one behavioral dimension to improve administrative efficiency; second, making relatively minor wording changes and reducing the length of the scale anchors to lessen the reading difficulty as well as the time required to complete the scales.

Combat Performance Prediction Scale

Forms A and B of the Combat Performance Prediction Scale were administered at only one post during the Batch B field testing. The scale was administered to peer and supervisor raters during the rating sessions, along with the Army-wide and MOS-specific rating scales.

No meaningful differences were found in means and standard deviations between supervisor and peer raters, or combat and noncombat MOS, or among the six scale dimensions. All of the means are slightly above the scale midpoint

of 7.5. A very low reliability of .21 was obtained for the total score on all 76 items when ratings were pooled across raters and MOS.

A set of 40 items was selected from the pool of 76 items on the basis of content domain (dimension) coverage and psychometric properties. Psychometric properties considered included reliability, item-dimension correlation, item-total correlation, and means and standard deviations across MOS and rater groups. Responses to the questions concerning rating confidence and item applicability were also considered.

Vast improvement in total score reliability (i.e., .21 to .56) resulted when the 40 best items from among the 76 were selected. Total scale coefficient alpha remained at .94. The 40-item scale was judged to have sufficiently good psychometric properties to justify its use for all MOS in the Concurrent Validation data collection.

Administrative/Archival Indicators

The Personnel File Information Form (a self-report of 201 File information) was administered at every field test site. Using the same form, project staff extracted the same information from each soldier's 201 File, thus making possible a comparison of the two approaches. A total of 505 cases were available for administrative measures analyses.

Self-Report vs. File Data

For the Number of Awards variable, there was perfect correspondence between the two sources. For the other measures, which showed varying levels of agreement, a greater percentage of soldiers were reporting more occurrences of administrative measures being received than were found in their 201 Files (e.g., see Tables 3-15 and 3-16).

This situation was not surprising in light of our earlier exploration of 201 Files. According to regulations, not all letters, certificates, Articles 15, etc. are placed in 201 Files, and some documents are removed after a certain period of time. Also, while 201 Files are the most timely official source of information, they are certainly not updated daily. Thus, discrepancies in the reported direction were not unexpected. If soldiers had reported more positive documents, such as letters and certificates, and fewer negative documents, such as Articles 15, when compared with the file data, then the self-report data would surely be suspect. However, soldiers reported receiving more negative as well as more positive documents.

Correlations were computed between the six administrative measures and Army-wide supervisor and peer ratings, respectively. Relationships obtained from the self-report approach were generally higher than those obtained from 201 Files.

To further investigate why self-report differed from file information, staff personnel conducted an outlier analysis by talking with individual soldiers, trying to determine the extent to which they were counting the items that we intended to be counted. If the soldier was interpreting the question as we intended, we then asked for possible explanations as to why a self-reported item was not found in the 201 File.

Table 3-15

Comparison of Letters/Certificates Information Obtained From
Self-Report and 201 Files: Batch A

Self-Report	201 File							Total
	0	1	2	3	4	5	6	
0	178	9	2	0	0	1	0	190
1	80	20	3	1	0	0	0	104
2	60	21	6	0	1	0	0	88
3	38	11	6	3	0	0	0	58
4	24	8	5	4	1	0	1	43
5	7	4	1	0	1	0	0	13
6	5	1	0	1	1	0	0	8
7	0	0	1	0	0	0	0	1
Total	392	74	24	9	4	1	1	505

Table 3-16

Comparison of Articles 15/FLAG Information Obtained From
Self-Report and 201 Files: Batch A

Self-Report	201 File				Total
	0	1	2	3	
0	320	10	2	0	332
1	73	6	4	0	83
2	38	13	2	1	54
3	18	8	1	0	27
4	2	1	1	1	5
5	1	1	0	0	2
6	1	0	0	0	1
7	1	0	0	0	1
Total	454	39	10	2	505

Some of the reasons confirmed earlier suspicions, such as "Counted
training certificates," "Counted certificate/letter that accompanied award,"
and "Recently received, paperwork not completed." Other reasons were
unexpected, such as "Counted Levy alert" as a FLAG action; a Levy alert is a
notification of an impending transfer. The lesson learned was a simple one:
For the Concurrent Validation data collection the self-report questions needed
to be more detailed, and even more clearly specified.

Revisions for Concurrent Validation

After the field tests of the Personnel File Information Form, it was concluded that self-report yields the most timely and complete data. However, a number of revisions were made in the self-report. The Military Training Courses variable was dropped from consideration because it had little variance and showed very low relationships with other measures. Further, since the earlier 201 File-EMF comparison showed almost perfect agreement for the Promotion Rate and Reenlistment Eligibility variables, and since monthly updates of the EMF have become available and there is no longer a need to collect this information from the field, the Reenlistment Eligibility question and three questions used to compute Promotion Rate were dropped from the Personnel File Information Form. Finally, as mentioned above, the remaining questions were made more detailed.

Training Achievement Tests

Training achievement data are for Batches A and B only (nine MOS). These data were collected both from trainees as they completed their respective course and from job incumbents during the Batch A and B field tests. Trainee and field test job incumbents results match; that is, coefficient alpha for both the trainee and job incumbent samples was .88. Mean correct for trainees was 53.9 percent, compared to 54.5 percent for job incumbents.

Reduction in Number of Items for Concurrent Validation

Because of time constraints, the length for the Concurrent Validation versions of the training tests would be limited to approximately 150 items. To reduce the size of the item pool, any items that had been rated not relevant to the job and also not relevant to training were dropped first. Next, items that had been rated lowest in importance and/or highest in difficulty were dropped. Because the training performance domain was assumed to be multidimensional, items were not usually eliminated solely because of a low correlation with the total test score. However, some items were dropped that exhibited the three characteristics of (a) low pass rate, (b) negative item-total correlation, and (c) a distractor or distractors with a high positive item-total r. During the revision of the item pools, the relative frequency of items in each job task duty area was maintained.

The numbers of items remaining on each test after the revisions had been made are reported in Tables 3-17, 3-18, and 3-19. The versions to be used for the Concurrent Validation contained the number of items shown in the columns on the far right. The tables for Batches A and B differ slightly from the table for Batch Z because many of the Batch A and B item reductions were made using field test data, which were not obtained for Batch Z. Before being administered to job incumbents as part of the Concurrent Validation, each item pool was submitted to the appropriate TRADOC Proponent for review. The number of items sent out for review and the number of items eliminated, added, or modified as a result of this review are also summarized.

Comparison of Initial and CV Item Pools

When initial item pool and Concurrent Validation versions are compared, there is a small increase in the percentage of items rated very important and

Table 9-17

Number of Items in Training Achievement Tests at Each Stage of Development: Batch A

MOS	Initial Item Pool No.[a]	No. of Cuts by Category		No. of Items Sent to Proponent Review[b]	Proponent Review		No. of Items Remaining
		Not Relevant	Low Importance or Poor Item Characteristics		Cut/Added	Modified	
13B (SP)	163/68[c]	18	75	138	2/0	0	136
13B (T)	163/67[c]	14	55	161	5/0	0	156
64C	228	2	86	140	12/0	70	128
71L	130	1	28	101	6/10	12	105
95B	223	11	72	140	6/5	9	139

[a] Items field tested.

[b] Reflects one or more Proponent reviews.

[c] There were 163 items common between the SP & T versions; 68 items were unique to SP and 67 unique to T.

94

Table 3-18

Number of Items in Training Achievement Tests at Each Stage of Development: Batch B

MOS	Item Pool No.		No. of Cuts by Category		No. of Items Sent to Proponent Review	Proponent Review			
	School	Field Test	Not Relevant	Low Importance or Poor Item Characteristics		Cut/ Added	Modified	Items Added to Rebalance Budget	No. of Items Remaining
11B	200	199	5	13	181	35/0	13	4	150
19E	214	202	2	21	179	17/0	8	0	162
31C	192	199	5	15	179	4/0	7	0	175
63B	238	217	2	47	168	24/23[a]	81	5	172
91A	299	260	5	46	209	34/0[b]	17	0	175

[a] Reflects two Proponent reviews.

[b] Reflects an additional cut made in conjunction with Proponent review to bring the test down to 175 items.

Table 3-19

Number of Items in Training Achievement Tests at Each Stage of Development: Batch 2

MOS	No. of Items Sent to Proponent Review	Cut/Added	Modified	Additional Cuts Based on Low Importance or Poor Item Characteristics	Items Added to Rebalance Budget	No. of Items Remaining
12B	211	0/0	5	35	0	176
16S	202	49/0	35	5	1	149
27E	205	2/0	9	28	0	175
51B	202	8/0	5	31	0	163[a]
54E	207	63/0	22	5	6	145
55B	212	0/0	5	31	0	181
67N	207	6/0	0	26	0	175
76W	195	1/0	0	19	0	175
76Y	188	2/0	11	11	0	175
94B	187	3/0	4	8	0	175

[a] Reduced to this number to eliminate time-consuming math items.

96

the combat scenario (Very Important, 33.1 to 34.0%; Of Little Importance, 22.8 to 20.6%) and the garrison scenario (Very Important, 43.1 to 46.5%; Of LittleImportance, 11.2 to 8.3%). These changes are all in the expected direction, given the procedures that were used to revise the initial item pools.

For the version of the tests administered as part of the Concurrent Validation, the distribution across relevance categories is nearly the same as for the original item pool.

Chapter 4
THE CONCURRENT VALIDATION

SAMPLES AND PROCEDURES

The nomenclature for MOS groupings was changed slightly for the Concurrent Validation, with previously designated Batch A and Batch B MOS becoming Batch A. The remaining 10 MOS were still designated as Batch Z, as listed in Table 4-1.

Table 4-1

MOS in the Concurrent Validation Phase of Project A

Batch A MOS		Batch Z MOS	
11B	Infantryman	12B	Combat Engineer
13B	Cannon Crewman	16S	MANPADS Crewman
19E	Armor Crewman	27E	TOW/Dragon Repairer
31C	Single Channel Radio Operator	51B	Carpentry/Masonry Specialist
63B	Light Wheel Vehicle Mechanic	54E	Chemical Operations Specialist
64C	Motor Transport Operator*	55B	Ammunition Specialist
71L	Administrative Specialist	67N	Utility Helicopter Repairer
91A	Medical Specialist	76W	Petroleum Supply Specialist
95B	Military Police	76Y	Unit Supply Specialist
		94B	Food Service Specialist

* In the latter part of the CV phase, MOS 64C became MOS 88M.

Collection of CV data was planned to begin in May 1985, using procedures that had been tried out and refined during the predictor and criterion field tests, and 13 data collection sites in the CONUS and sites in USAREUR. Data collection actually began 10 June 1985 and was concluded 13 November 1985. The data were collected by on-site teams made up of seven or eight project staff members. At the peak of data collection, seven teams (one per post) were operating.

Samples Obtained

The final sample sizes obtained are shown by post and by MOS in Table 4-2. A target sample size of 600-700 job incumbents per MOS was the overall goal, but in some MOS, the sample was smaller, either because the MOS simply is not that large or because not enough incumbents with the appropriate accession dates were available at the various sites.

Table 4-2

Concurrent Validation Sample Soldiers by MOS by Location

Location	Batch 1 MOS									Batch 2 MOS										Total	% Total
	11B	13B	19E	31C	63B	64C	71L	91A	95B	12B	16S	27E	51B	54E	55B	67N	76W	76Y	94B		
Fort Benning	45	23	41	7	13	39	16	9	13	13	15	3	0	12	18	9	13	15	12	316	3.35
Fort Bliss	0	20	30	15	61	45	17	0	44	15	5	2	0	14	0	12	6	31	30	347	3.68
Fort Bragg	68	46	0	0	37	25	41	10	72	82	75	13	19	72	20	7	42	39	62	730	7.74
Fort Campbell	90	28	0	20	60	45	54	44	63	90	23	10	0	32	18	42	51	61	46	757	8.03
Fort Carson	60	50	77	30	49	53	30	33	46	49	57	13	0	25	7	0	23	40	47	689	7.31
Fort Hood	26	56	0	30	40	28	38	50	60	51	60	4	12	62	36	44	72	41	57	767	8.13
Fort Knox	29	32	121	16	38	48	22	45	31	43	10	6	0	8	12	0	10	29	34	524	5.56
Fort Lewis	75	46	13	11	43	46	23	27	56	27	25	1	11	51	31	20	48	41	36	631	6.69
Fort Ord	38	0	0	14	30	42	31	63	51	51	7	8	1	4	7	15	23	40	28	425	4.51
Fort Polk	73	47	19	29	47	47	18	46	44	60	45	9	8	16	7	23	26	51	35	648	6.87
Fort Riley	30	43	55	27	26	45	35	30	40	31	20	8	8	25	52	0	20	35	45	579	6.14
Fort Sill	0	108	0	20	43	51	44	0	29	42	11	0	0	0	0	15	7	35	32	437	4.63
Fort Stewart	44	46	39	17	28	51	31	45	45	30	39	9	8	17	29	26	44	34	33	617	6.54
UNKNOWN	127	122	120	130	122	121	114	119	118	120	78	61	41	96	54	63	105	134	113	1953	20.80
Total	702	667	553	366	637	686	514	501	692	704	470	167	108	434	291	276	490	630	612	9430	
% Total	7.44	7.07	5.33	3.88	6.76	7.27	5.45	5.31	7.34	7.47	4.90	1.56	1.15	6.60	3.09	2.93	5.20	6.68	6.49		

Predictor and Criterion Measures

The full array of predictor and criterion measures used in the Concurrent Validation is described at some length in the FY86 Annual Report (Campbell, 1987) and in the development and field test reports for each major type of instrument. The variables in each domain are listed in Tables 4-3 and 4-4. In the Concurrent Validation one-half day was devoted to predictor measurement and one and one-half days to criterion measurement.

While the same predictor battery was used for all the MOS, the criterion measures used for Batch A MOS were different than those used for MOS in Batch Z. The major distinction is that the MOS-specific job performance and job knowledge measures were not developed for the 10 MOS in Batch Z. For these jobs only Army-wide measures and the training achievement tests were administered.

Data Collection Team Composition and Training

Each data collection team was composed of a test site manager and six or seven project staff members who were responsible for administering tests and rating scales. The teams were made up of a combination of regular project staff and individuals (e.g., graduate students) specifically hired for the data collection effort. The test site manager had participated extensively in the field tests. The team was assisted by eight NCO scorers (for the hands-on tests), one company-grade officer POC, and up to five NCO support personnel, all provided by the post. The project data collection teams were given 3 days of training at a central location (Alexandria, VA). The eight NCO scorers who were required to administer and score the hands-on tests were recruited and trained at each post, using procedures very similar to those used in the criterion field tests. Training required one full day during which scorers had the opportunity to take the tests themselves and undergo multiple practice trials in scoring each task, with feedback from the project staff.

Concurrent Validation Analyses

The basic analytic steps for the Concurrent Validation data were as outlined below. The overall goal was to move systematically from the raw data, which consist of thousands of elements of information on each individual, to estimates of selection validity, differential validity, and selection/classification utility.

General Steps

The general steps in the analysis were as follows:

(1) Prepare and edit individual data files.

(2) Determine basic scores for the predictor variables.

(3) Determine basic scores for the criterion variables.

(4) Describe the latent structure of the predictor and criterion covariance matrixes.

101

Table 4-3

Summary of Predictor Measures Used in Concurrent Validation:
The Trial Battery

Name	Number of Items
COGNITIVE PAPER-AND-PENCIL TESTS	
Reasoning Test (Induction-Figural Reasoning)	30
Object Rotation Test (Spatial Visualization-Rotation)	90
Orientation Test (Spatial Orientation)	24
Maze Test (Spatial Orientation)	24
Map Test (Spatial Orientation)	20
Assembling Objects Test (Spatial Visualization-Rotation)	32
COMPUTER-ADMINISTERED TESTS	
Simple Reaction Time (Processing efficiency)	15
Choice Reaction Time (Processing efficiency)	30
Memory Test (Short-term memory)	36
Target Tracking Test 1 (Psychomotor precision)	18
Perceptual Speed and Accuracy Test (Perceptual speed and accuracy)	36
Target Tracking Test 2 (Two-hand coordination)	18
Number Memory Test (Number Operations)	28
Cannon Shoot Test (Movement judgment)	36
Identification Test (Perceptual speed and accuracy)	36
Target Shoot Test (Psychomotor precision)	30
NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES	
Assessment of Background and Life Experiences (ABLE)	209
Adjustment	
Dependability	
Achievement	
Physical Condition	
Leadership	
Locus of Control	
Agreeableness/Likability	
Army Vocational Interest Career Examination (AVOICE)	176
Realistic Interests	
Conventional Interests	
Social Interests	
Enterprising Interests	
Artistic Interests	

Table 4-4

Summary of Criterion Measures Used in Batch A and Batch Z
Concurrent Validation Samples

<u>Performance Measures Common to Batch A and Batch Z</u>

- Army-wide rating scales (all obtained from both supervisors and peers).

    - Ten behaviorally anchored rating scales (BARS) designed to measure factors of non-job-specific performance.

    - Single scale rating of overall effectiveness.

    - Single scale rating of NCO potential.

- Combat Prediction scale containing 40 items.

- Paper-and-pencil tests of training achievement developed for each of the 19 MOS (130-210 items each).

- Personnel File information form developed to gather objective archival records data (awards and letters, rifle marksmanship scores, physical training scores, etc.).

<u>Performance Measures for Batch A Only</u>

- Job sample (hands-on) tests of MOS-specific task proficiency.

    - Individual is tested on each of 15 major job tasks in an MOS.

- Paper-and-pencil job knowledge tests designed to measure task-specific job knowledge.

    - Individual is scored on 150 to 200 multiple-choice items representing 30 major job tasks. Ten to 15 of the tasks were also measured hands-on.

- Rating scale measures of specific task performance on the 15 tasks also measured with the knowledge tests. Most of the rated tasks were also included in the hands-on measures.

- MOS-specific behaviorally anchored rating scales (BARS). From six to 12 BARS were developed for each MOS to represent the major factors that constitute job-specific technical and task proficiency.

<u>Performance Measures for Batch Z Only</u>

- Additional Army-wide rating scales (all obtained from both supervisors and peers).

    - Ratings of performance on 11 common tasks (e.g., basic first aid).

    - Single scale rating on performance of specific job duties.

<u>Auxiliary Measures Included in Criterion Battery</u>

- A Job History Questionnaire which asks for information about frequency and recency of performance of the MOS-specific tasks.

- Army Work Environment Questionnaire - 53 items assessing situational/environmental characteristics, plus 46 items dealing with leadership.

- Measurement Method Rating obtained from all participants at the end of the final testing session.

---

(5) Determine how well each predictor construct predicts each criterion factor (for each MOS).

(6) Determine incremental validities (if any) of new predictors over ASVAB for each criterion factor within each MOS.

Missing Values

Because extensive multivariate analyses requiring complete data were to be performed, the treatment of missing values was an important concern (Young, Harris, Hoffman, Houston, & Wise, 1987). Cases with significant amounts of missing data (10% for written tests, 15% for hands-on tests and rating scales) were dropped from the analysis of that instrument. In cases where lesser amounts of data were missing, either examinee means or variable means were substituted for missing values. For these data, the PROC IMPUTE statistical procedure was used to derive proxy values for missing scale scores, and for missing step scores in the hands-on analyses. These procedures enabled retention of 90-95 percent of the soldiers in each MOS.

The PROC IMPUTE procedure essentially substitutes for the missing variable a value observed for a respondent who is very similar to the examinee. This procedure has been shown to be significantly better than ordinary least squares (OLS) regression procedures (e.g., BMDPAM) in reproducing correlation and variance estimates, as the regression approaches tend to underestimate variances and to spuriously inflate correlations.

Predictor Score Analyses

After data preparation, basic item analyses, and the initial score generation, the principal objectives for the predictor analyses were to generate the basic summary scores that would enter the initial prediction equation for each MOS. The basic steps were as follows:

(1) Using the initial scores, conduct item/scale score analyses.

(2) Compute scale reliabilities and descriptive statistics.

(3) Develop predictor construct scores via factor analysis.

(4) Estimate predictor factor (construct) scores via a simple weighted sum.

Criterion Score Analyses

After data preparation had been completed, the objectives for the criterion analyses were to identify an array of basic criterion variables (i.e., scores), investigate the latent structure of those variables, and determine the principal criterion component scores.

Predictor/Criterion Interrelationships

After the above steps were carried out, the basic variables and the best-fitting model for both the predictors and the performance measures had been identified. They provided the variables to be used for establishing the selection/classification validity of the new predictor battery and for

determining differential validity across criterion constructs, across jobs, and across subgroups.

DEVELOPMENT OF PREDICTOR SCORES AND COMPOSITES

Basic Predictor Scores for the Trial Battery

A total of 69 scores were generated from the Trial Battery. Forty-three came from the non-cognitive inventories--Assessment of Background and Life Experiences (ABLE), the Army Vocational Interest Career Examination (AVOICE), and the Job Orientation Blank (JOB), which had been included in the AVOICE for the Trial Battery. Six scores came from the six paper-and-pencil cognitive tests. For the computer-administered tests, a number of alternative methods of scoring, such as slopes, intercepts, and different methods of computing means (e.g., different procedures for trimming items before computing means), were evaluated. Generally speaking, the computerized test scores selected for additional analyses were those that were most reliable and could be interpreted in a straightforward way.

The Ns, means, standard deviations, reliabilities, and uniqueness (from ASVAB) coefficients for scores on the cognitive paper-and-pencil tests are shown in Table 4-5. Similar data are shown in Tables 4-6 and 4-7 for the computer-administered tests, and in Tables 4-8, 4-9, and 4-10 for the ABLE, AVOICE, and JOB scale scores. Uniqueness coefficients are not shown for these instruments, but range from .40 to .88, with a median u^2 of .79 for ABLE, .80 for AVOICE, and .57 for JOB.

In general, the battery exhibited quite good psychometric properties, with the exception of low reliabilities on some computer-administered test scores. The low reliabilities tended to be characteristic of the proportion correct scores, which was expected. That is, the items can almost always be answered correctly if the examinee takes enough time, which restricts the range on the proportion correct scores. However, it increases the variance (and reliability) on the decision time scores.

Formation of Predictor Composites

Preliminary analyses of the Trial Battery predictor tests indicated that reliable predictor scores could be computed from the six spatial tests (i.e., the paper-and-pencil cognitive tests), the 10 computerized tests, and the temperament, vocational interest, and job reward inventories (Peterson, et al., 1987). In addition, scores from the nine ASVAB subtests were available from Army records. Table 4-11 shows how these predictor scores were distributed among various domains within the predictor space. The ASVAB subtests measured nine cognitive abilities. The paper-and-pencil cognitive tests measured six different aspects of spatial ability. The 10 computerized tests yielded 20 measures of perceptual-psychomotor abilities. The ABLE provided measures of 11 temperament/ biographical traits. The AVOICE assessed 22 vocational interests. Finally, the JOB measured six types of job reward preferences.

Table 4-5

Concurrent Validity Data Analysis: Statistics for Paper-and-Pencil
Cognitive Tests

Test	N	Mean	SD	Split-Half Reli-ability[a]	Test-Retest Reli-ability[b]	Uniqueness Estimate
Assembling Objects	9,343	23.3	6.71	.91	.70	.65
Object Rotation	9,345	62.4	19.06	.99	.72	.81
Maze	9,344	16.4	4.77	.96	.70	.74
Orientation	9,341	11.0	6.18	.89	.70	.60
Map	9,343	7.7	5.51	.90	.78	.46
Reasoning	9,332	19.1	5.67	.87	.65	.54

[a]Split-half reliability estimates were calculated using the odd-even
procedure with the Spearman-Brown correction for test length.

[b]Test-retest reliability estimates are based on a sample of 468 to 487
subjects. The test-retest interval was 2 weeks.

Because of multicollinearity and the ratio of number of variables to
sample size, 78 separate predictor scores were too many to retain. Conse-
quently, the 78 predictor test and scale scores were combined into 24 predic-
tor composites before predictor-criterion relationships were computed. With
one exception (which will be noted), these composites were formed simply by
summing standardized test or scale scores.

Table 4-6

Concurrent Validity Data Analysis: Statistics for Computerized
Psychomotor Tests

Test	N	Mean	SD	Odd-Even Reliability[a]	Test-Retest Reliability[b]	Uniqueness Estimate
Target Tracking 1						
Mean Log (Distance + 1)	9,251	2.98	.49	.98	.74	.82
Target Tracking 2						
Mean Log (Distance + 1)	9,239	3.70	.51	.98	.85	.79
Target Shoot						
Mean Log (Distance + 1)	8,892	2.17	.24	.74	.37	.70
Mean Time to Fire	8,892	235.39	47.78	.85	.58	.78
Cannon Shoot						
Mean Absolute Time Discrepancy	9,234	43.94	9.57	.65	.52	.56

[a]Time-to-fire and time-discrepancy measures are in hundredths of seconds.
 Logs are natural logs.

[b]Test-retest reliability estimates are based on sample sizes of 468 to 487.
 The test-retest interval was 2 weeks.

Table 4-7

Concurrent Validity Data Analysis: Statistics for Computerized
Perceptual Tests

Test & N for each		Mean	SD	Odd-Even Reliability[a]	Test-Retest Reliability[b]	Uniqueness Esitmate
Simple Reaction Time (SRT)						
Decision Time Mean	9,255	31.84	14.82	.88	.23	.87
Proportion Correct	9,255	.98	.04	.46	.02	.44
Choice Reaction Time (CRT)						
Decision Time Mean	9,269	40.93	9.77	.97	.69	.93
Proportion Correct	9,269	.98	.03	.57	.23	.55
Short Term Memory (STM)						
Decision Time Mean	9,149	87.72	24.03	.96	.66	.93
Proportion Correct	9,149	.89	.08	.60	.41	.55
Perceptual Speed & Accuracy (PSA)						
Decision Time Mean	9,244	236.91	63.38	.94	.63	.92
Proportion Correct	9,244	.87	.08	.65	.51	.61
Target Identification (TID)						
Decision Time Mean	9,105	193.65	63.13	.97	.78	.83
Proportion Correct	9,105	.91	.07	.62	.40	.59
Number Memory						
Final Response Time Mean	9,099	160.70	42.63	.88	.62	.67
Input Response Time Mean	9,099	142.84	55.24	.95	.47	.85
Operations Response Time Mean[c]	9,099	233.10	79.72	.93	.73	.66
Proportion Correct	9,099	.90	.09	.59	.53	.39
SRT-CRT-STM-PSA-TID						
Pooled Mean Movement Time[c]	8,962	33.61	8.03	.74	.66	.71

[a]Times are given in hundredths of seconds.
[b]N = 460-479 for test-retest correlations. The test-retest interval was 2 weeks.
[c]Coefficient Alpha reliability estimates.

Table 4-8

ABLE Scale Statistics for Total Group[a]: Trial Battery

ABLE Scale	No. Items	N	Mean	SD	Median Item- Total Corre- lation	Internal Consis- tency Relia- bility (Alpha)	Test- Retest Relia- bility[b]
Substantive Scales							
Emotional Stability	17	8,522	39.0	5.45	.39	.81	.74
Self-Esteem	12	8,472	28.4	3.70	.39	.74	.78
Cooperativeness	18	8,494	41.9	5.28	.39	.81	.76
Conscientiousness	15	8,504	35.1	4.31	.34	.72	.74
Nondelinquency	20	8,482	44.2	5.91	.36	.81	.80
Traditional Values	11	8,461	26.6	3.72	.36	.69	.74
Work Orientation	19	8,498	42.9	6.06	.41	.84	.78
Internal Control	16	8,485	38.0	5.11	.39	.78	.69
Energy Level	21	8,488	48.4	5.97	.38	.82	.78
Dominance	12	8,477	27.0	4.28	.44	.80	.79
Physical Condition	6	8,500	14.0	3.04	.60	.84	.85
Response Validity Scales							
Unlikely Virtues	11	8,511	15.5	3.04	.34	.63	.63
Self-Knowledge	11	8,508	25.4	3.33	.36	.65	.64
Non-Random Response	8	8,559	7.7	.59			.30
Poor Impression	23	8,492	1.5	1.85	.20	.63	.61

[a]Total group after screening for missing data and random responding.

[b]N = 408-414 for test-retest correlation. Test-retest interval was 2 weeks.

Table 4-9

AVOICE Scale Statistics for Total Group[a]: Trial Battery

AVOICE Scale	No. Items	N	Mean	SD	Median Item-Total Corre-lation	Internal Consis-tency Relia-bility (Alpha)	Test-Retest Relia-bility[b]
Clerical/ Administrative	14	8,463	39.6	10.81	.67	.92	.78
Mechanics	10	8,382	32.1	9.42	.80	.94	.82
Heavy Construction	13	8,488	39.3	10.54	.68	.92	.84
Electronics	12	8,359	38.4	10.22	.70	.94	.81
Combat	10	8,466	26.5	8.35	.65	.90	.73
Medical Services	12	8,364	36.9	9.54	.68	.92	.78
Rugged Individualism	15	8,396	53.3	11.44	.58	.90	.81
Leadership/Guidance	12	8,444	40.1	8.63	.62	.89	.72
Law Enforcement	8	8,471	24.7	7.37	.65	.89	.84
Food Service - Professional	8	8,472	20.2	6.50	.67	.89	.75
Firearms Enthusiast	7	8,397	23.0	6.36	.66	.89	.80
Science/Chemical	6	8,468	16.9	5.33	.70	.85	.74
Drafting	6	8,493	19.4	4.97	.66	.84	.74
Audiographics	5	8,473	17.6	4.09	.69	.83	.75
Aesthetics	5	8,413	14.2	4.13	.59	.69	.73
Data Processing	4	8,224	14.0	3.99	.78	.90	.70
Food Service - Employee	3	8,304	5.1	2.08	.54	.73	.56
Mathematics	3	8,421	9.6	3.09	.78	.88	.75
Electronic Communications	6	8,403	18.4	4.66	.60	.83	.68
Warehousing/Shipping	2	8,407	5.8	1.75	.44	.61	.54
Fire Protection	2	8,431	6.1	1.96	.62	.76	.67
Vehicle/Equipment Operator	3	8,378	8.8	2.65	.51	.70	.68

[a]Total group after screening for missing data and random responding.

[b]N = 389-409 for test-retest correlation. Test-retest interval was 2 weeks.

Table 4-10

JOB Scale Statistics for Total Group[a]: Trial Battery

JOB	No. Items	N	Mean	SD	Median Item-Total Correlation	Internal Consistency Reliability (Alpha)
Job Security	10	7,809	43.6	4.51	.54	.84
Job Pride	5	7.817	21.6	2.33	.43	.57
Serving Others	3	7,784	12.1	1.83	.52	.66
Autonomy	4	7,817	15.1	2.29	.31	.50
Routine	4	7,707	9.6	2.30	.25	.46
Ambition	3	7,751	12.4	1.63	.35	.49

[a]Total group after screening for missing data and random responding.

Table 4-11

Assessment of the Selected Measures with Reference to the Predictor Space

Predictor Domain	Measures[a]	Number of Test or Scale Scores	Number of Composite Scores
General Cognitive Ability	Armed Services Vocational Aptitude Battery (ASVAB)	9 Subtests	4
Spatial Ability	Spatial Test Battery	6 Tests	1
Perceptual-Psychomotor Abilities	Computerized Battery	20 Tests	6
Temperament	Assessment of Background and Life Experiences (ABLE)	11 Scales[b]	4
Vocational Interests	Army Vocational Interest Career Examination (AVOICE)	22 Scales	6
Job Reward Preferences	Job Orientation Blank (JOB)	6 Scales	3

[a]All measures except the ASVAB were developed specifically for Project A.
[b]The ABLE included four additional response validity scales.

Three goals guided the formation of composite scores. First, there was an attempt to keep the number of composites to a minimum. Second, homogeneity within composites was maximized. Third, even if two or more test or scale scores were reasonably highly correlated and had similar patterns of factor loadings, they were grouped into the same composite only if they were expected to have similar patterns of correlations with job performance.

Figure 4-1 shows how the nine ASVAB subtests were combined into four composite scores: Technical, Quantitative, Verbal, and Speed. In computing the Technical composite score, the Electronics Information subtest received a weight of one-half unit while the Mechanical Comprehension and Auto-Shop subtests received unit weights, because a factor analysis indicated that the loading of the Electronics Information subtest on the Technical factor of the ASVAB was only about one-half as large as the loading of the Mechanical Comprehension and Auto-Shop subtests.

The six spatial tests were all highly intercorrelated and as Figure 4-2 shows, were combined into a single composite score. Six composite scores were computed from the 20 perceptual-psychomotor test scores from the computerized battery (Figure 4-3). Four temperament composites were computed from the ABLE scales (see Figure 4-4) and six vocational interest composites were computed from the 21 AVOICE scales (see Figure 4-5). Finally, the six scales of the JOB were combined into three composites (Figure 4-6).

All subsequent predictor validation analyses were based on these 24 basic scores. They are portrayed in summary form in Table 4-12. The tests and inventory scales from the Trial Battery which were used to form simple sum factor scores are listed under each factor title.

DEVELOPMENT OF BASIC JOB PERFORMANCE CRITERION SCORES

During the Concurrent Validation, Project A collected 12 hours of criterion data from 5,000 incumbents in nine MOS (Batch A) and 4 hours of data from 4,500 incumbents in 10 MOS (Batch Z). For each individual in Batch A there were approximately 350 knowledge test items, 15 hands-on task scores, 95 rating scales from each of three raters, and 6 administrative indexes. The first major step in reducing these multiple bits of information to scores on the major components of performance was the development of the "basic" criterion scores that could be used in covariance analyses of the latent structure. The procedures that the project staff used to obtain these basic scores are summarized below.

Criterion Scores for the Hands-On and Knowledge Tests

To reduce the number of criterion scores derived from the hands-on tests and job knowledge tests, the task domains for each of the nine Batch A MOS were reviewed by project staff and tasks were clustered into a set of functional categories on the basis of task content. Ten of the categories applied to all MOS and consisted primarily of common tasks. In addition, each MOS,

112

```
┌─────────────────────────────┐                    ┌──────────────┐
│ Mechanical Comprehension    │                    │              │
│ Auto-Shop Information       │ ─────────────────▶ │  Technical   │
│ Electronics Information     │                    │              │
└─────────────────────────────┘                    └──────────────┘


┌─────────────────────────────┐                    ┌──────────────┐
│ Math Knowledge              │                    │              │
│ Arithmetic Reasoning        │ ─────────────────▶ │ Quantitative │
│                             │                    │              │
└─────────────────────────────┘                    └──────────────┘


┌─────────────────────────────┐                    ┌──────────────┐
│ Verbal                      │                    │              │
│ General Science             │ ─────────────────▶ │   Verbal     │
│                             │                    │              │
└─────────────────────────────┘                    └──────────────┘


┌─────────────────────────────┐                    ┌──────────────┐
│ Coding Speed                │                    │              │
│ Number Operations           │ ─────────────────▶ │   Speed      │
│                             │                    │              │
└─────────────────────────────┘                    └──────────────┘
```
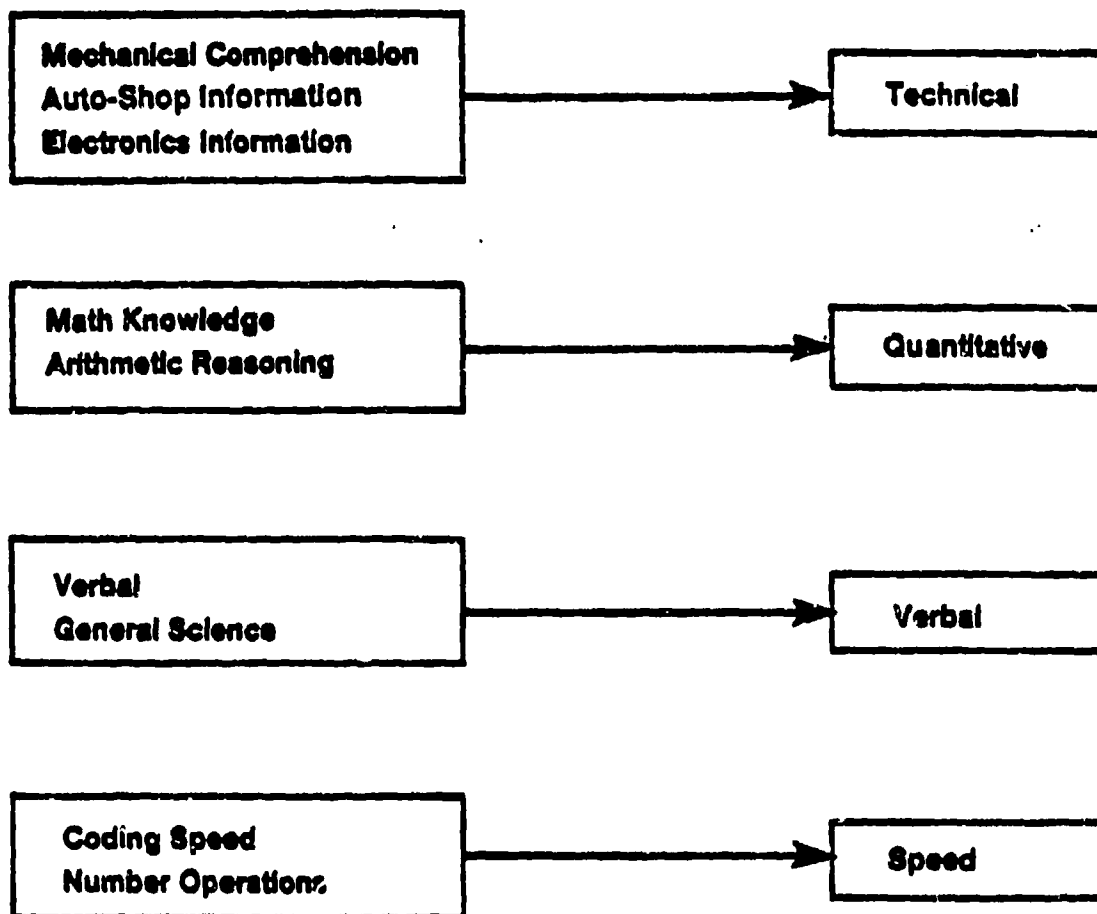
Figure 4-1. Formation of general cognitive ability composites
from ASVAB subtests.

113
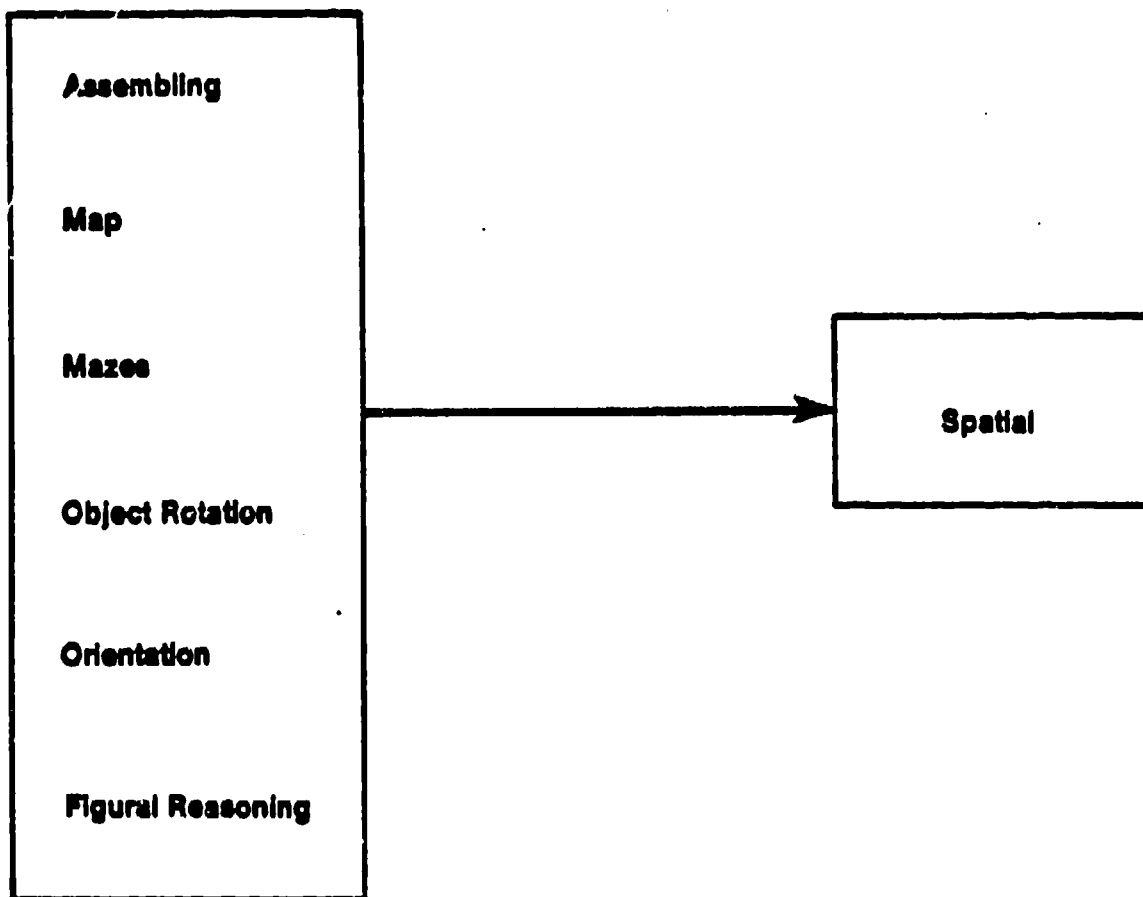
Figure 4-2. Formation of spatial ability composite from spatial battery test scores.

114

Cannon Shoot Test (Time Score)	
Target Shoot Test (Time to Fire)	
Target Shoot Test (Log Distance)	Psychomotor
Target Tracking 1 (Log Distance)	
Target Tracking 2 (Log Distance)	
Pooled Mean Movement Time	

Short-Term Memory Test (Decision Time)	Complex
Perceptual Speed & Accuracy Test (Decision Time)	Perceptual
Target Identification Test (Decision Time)	Speed

Short-Term Memory Test (Percent Correct)	Complex
Perceptual Speed & Accuracy Test (Percent Correct)	Perceptual
Target Identification Test (Percent Correct)	Accuracy

Number Memory Test (Percent Correct)	
Number Memory Test (Initial Response Time)	Number
Number Memory Test (Mean Operations Response Time)	Speed and
Number Memory Test (Final Response Time)	Accuracy

| Choice Reaction Time (Decision Time) | Simple |
| Simple Reaction Time (Decision Time) | Reaction Speed |

| Choice Reaction Time (Percent Correct) | Simple |
| Simple Reaction Time (Percent Correct) | Reaction Accuracy |

NOTE: One computer test score, Choice Reaction Time (Decision Time Minus Simple Reaction Time), was not used in computing composite scores.

Figure 4-3. Formation of perceptual-psychomotor ability composites from computerized battery test scores.

115

Self-Esteem **Work Orientation** **Energy Level** →	**Achievement Orientation**
Conscientiousness **Non-Delinquency** →	**Dependability**
Emotional Stability →	**Adjustment**
Physical Condition →	**Physical Condition**
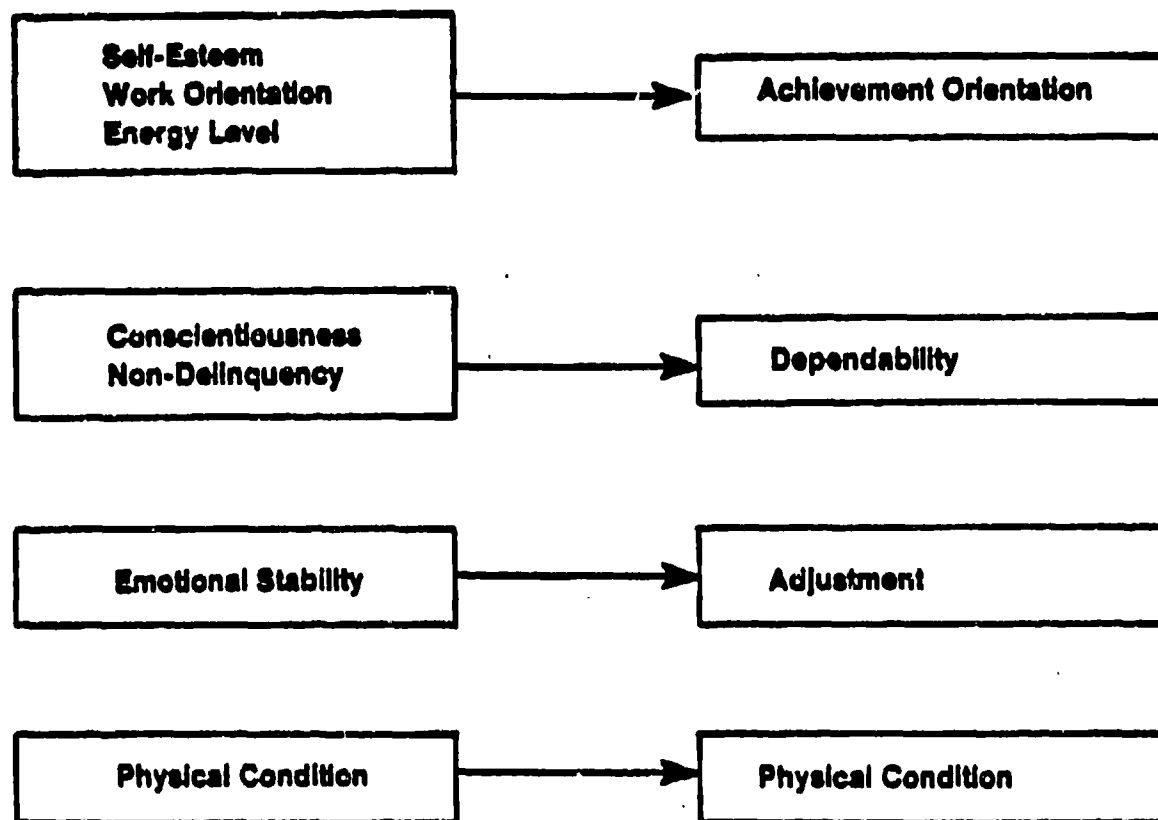
NOTE: Four ABLE scales (Dominance, Traditional Values, Cooperativeness, and Internal Control) were not used in computing composite scores.

Figure 4-4. Formation of temperament composites from ABLE scale scores.

Clerical/Administrative Medical Services Leadership/Guidance Science/Chemical Data Processing Mathematics Electronic Communications	→	Skilled Technical
Mechanics Heavy Construction Electronics Vehicle/Equipment Operator	→	Structural/Machine
Combat Rugged Individualism Firearms Enthusiast	→	Combat-Related
Drafting Audiographics Aesthetics	→	Audiovisual Arts
Food Service Professional Food Service Employee	→	Food Service
Law Enforcement Fire Protection	→	Protective Services

Figure 4-5. Formation of vocational interest composites from AVOICE scale scores.

117

```
┌─────────────────────┐                              ┌─────────────────────┐
│ Job Pride           │                              │ Organizational and  │
│ Job Security        │─────────────────────────────▶│ Co-Worker Support   │
│ Serving Others      │                              │                     │
│ Ambition            │                              │                     │
└─────────────────────┘                              └─────────────────────┘


┌─────────────────────┐                              ┌─────────────────────┐
│ Routine             │─────────────────────────────▶│ Routine Work        │
└─────────────────────┘                              └─────────────────────┘


┌─────────────────────┐                              ┌─────────────────────┐
│ Autonomy            │─────────────────────────────▶│ Job Autonomy        │
└─────────────────────┘                              └─────────────────────┘
```
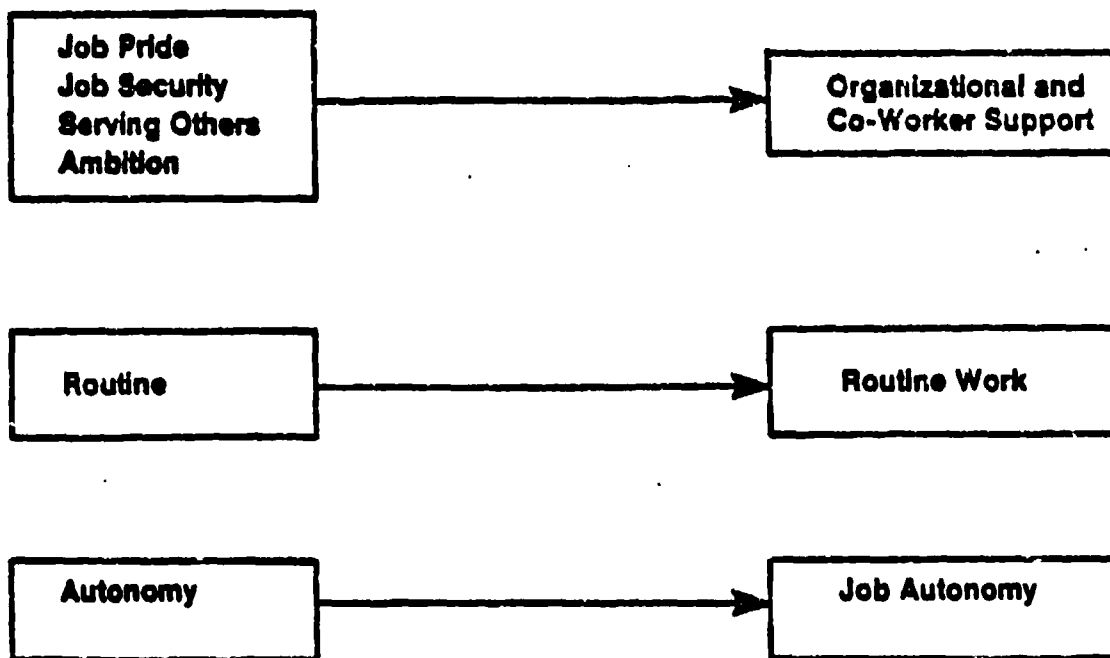
Figure 4-6. Formation of job reward preference composites from JOB scale scores.

Table 4-12

Ability, Temperament, and Interest Factors Identified via Analysis of the Concurrent Validation Data on 9,430 MOS Incumbents

FROM ASVAB SUBTESTS

Technical Factor
  Mechanical Comprehension
  Auto-Shop Information
  Electronics Information
Quantitative Factor
  Math Knowledge
  Arithmetic Reasoning
Verbal Factor
  Verbal
  General Science
Speed Factor
  Coding Speed
  Number Operations

FROM PAPER-AND-PENCIL TESTS

Overall Spatial Factor
  Assembling Objects Test
  Map Test
  Maze Test
  Object Rotation Test
  Orientation Test
  Figural Reasoning Test

FROM COMPUTERIZED MEASURES

Psychomotor Factor
  Cannon Shoot Test (Time score)
  Target Shoot Test (Time to fire)
  Target Shoot Test (Log distance)
  Target Tracking 1 (Log distance)
  Target Tracking 2 (Log distance)
  Pooled Mean Movement Time

Perceptual Speed Factor
  Short-Term Memory Test (Decision time)
  Perceptual Speed & Accuracy Test (Decision time)
  Target Identification Test (Decision time)

Perceptual Accuracy Factor
  Short-Term Memory Test (Percent correct)
  Perceptual Speed & Accuracy Test (Percent correct)
  Target Identification Test (Percent correct)

Number Speed/Accuracy Factor
  Number Memory Test (Percent correct)
  Number Memory Test (Initial decision time)
  Number Memory Test (Mean operations time)
  Number Memory Test (Final decision time)

Simple Reaction Speed Factor
  Choice Reaction Time (Decision time)
  Simple Reaction Time (Decision time)

Simple Reaction Accuracy Factor
  Choice Reaction Time (Percent correct)
  Simple Reaction Time (Percent correct)

FROM NON-COGNITIVE INVENTORIES

Achievement Factor
  Self-Esteem scale
  Work Orientation scale
  Energy Level scale

Dependability Factor
  Conscientiousness scale
  Non-delinquency scale

Adjustment Factor
  Emotional Stability scale

Physical Condition Factor
  Physical Condition scale

Skilled Technical Interest Factor
  Clerical/Administrative
  Medical Services
  Leadership/Guidance
  Science/Chemical
  Data Processing
  Mathematics
  Electronic Communications

Structural/Machines Interest Factor
  Mechanics
  Heavy Construction
  Electronics
  Vehicle/Equipment Operator

Combat-Related Interest Factor
  Combat
  Rugged Individualism
  Firearms Enthusiast

Audiovisual Arts Interest Factor
  Drafting
  Audiographics
  Aesthetics

Food Service Interest Factor
  Food Service Professional
  Food Service Employee

Protective Services Interest Factor
  Law Enforcement
  Fire Protection

Preference for Organizational and Co-worker Support
  Job Pride
  Job Security
  Serving Others
  Ambition

Preference for Routine Work
  Routine

Preference for Job Autonomy
  Autonomy

except for 11B (Infantryman) and 64C (Motor Transport Operator), had two to five MOS-specific categories. The ten common categories were sufficient to account for all tasks in 11B and 64C.

After category definitions had been written, three members of the project staff independently classified the 30 tasks in each MOS into one of the ten common categories or into an MOS-specific category. The level of perfect agreement in the assignment of tasks to categories was over 90 percent in every MOS. These same functional categories were used by the project staff to sort the school knowledge test items. The titles of the functional category definitions are presented in Figure 4-7.

Scores for the functional categories were computed by taking the sum of the hands-on task test steps (adjusted for length) or job knowledge test items in each category.

Separate principal components analyses were then carried out for each MOS, using the functional category score intercorrelation matrix as the input. The results of factor analyses performed in each of the nine MOS suggested a similar set of category clusters, with minor differences, across all nine MOS. The ten functional categories that cut across MOS and the several technical functional categories that were unique to particular MOS were reduced to six basic scores:

(1)   Communications - including the Communications functional category.

(2)   Vehicles - including the Vehicle Operation functional category, and for MOS 63B only the Vehicle Operation and Recovery category; for MOS 64C, the Vehicle Operation functional category went into the Technical cluster.

(3)   Basic Soldiering - including the Navigate, Weapons, Field Techniques, Customs and Laws, and Anti-Air/Tank Weapons categories.

(4)   Identify Targets - including the Identify Targets functional category.

(5)   Safety/Survival - including the First Aid and NBC functional categories.

(6)   Technical - including the functional categories peculiar to each MOS, comprising (usually) MOS-specific tasks; for MOS 64C, this cluster included the Vehicle Operation category, which comprises tasks central to the 64C job.

Although this set of clusters was not reproduced precisely for every one of the MOS, it appeared to be a reasonable portrayal of the nine jobs when a common set of clusters was imposed on all. Tables 4-13 and 4-14 show the range of correlations among the clusters and between the categories and the clusters, across the nine MOS.

120

Common Categories	MOS-Specific Categories
First Aid	**13B - Cannon Crewman**
NBC	Prepare, Operate, Maintain
Weapons	Howitzer and Ammunition
Navigate	Operate Howitzer Sights and
Field Techniques	Alignment Devices
Customs and Laws	
Communications	**19E - Tank Crewman**
Identify Targets	Operate Tanks
Anti-Air/Tank Weapons	Tank Gunnery
Vehicle Operation	
	31C - Single Channel Radio Operator
	Generators
	TTY Station and Net Operations
	Maintain TTY Electronic Equipment
	Operate TTY Electronic Equipment
	Install TTY Electronic Equipment
	63B - Light Wheel Vehicle Mechanic
	Electrical System
	Fuel/Cooling/Lubricating
	Brake/Steering/Suspension Systems
	Vehicle Operation and Recovery
	71L - Administrative Specialist
	Forms/Files Management
	Supervision/Coordination
	Correspondence
	Classified Material
	91A - Medical Specialist
	Clinic/Ward Treatment and Care
	Clinic/Ward Housekeeping
	Clinic/Ward Management
	95B - Military Police
	Responding to Alarms
	Patrol Duties
	Conduct MP Procedures
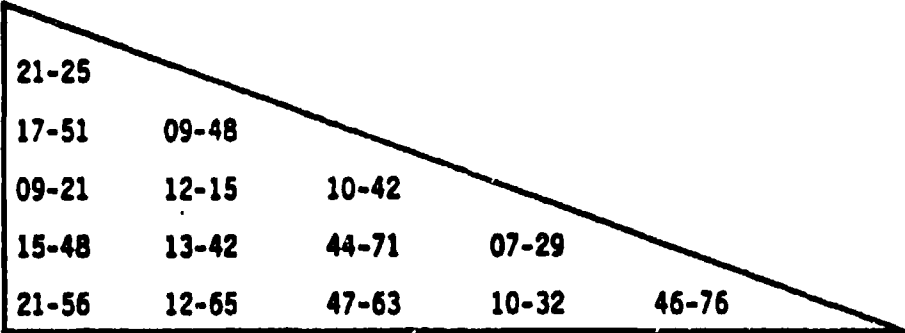
Figure 4-7. Functional task categories.

Table 4-13

Correlations Between Criterion Factor Scores and Functional Categories for Job Knowledge Component

FACTORS	Commo.	Vehicle	Basic	Identify	Survival	Technical
Communications						
Vehicles	21-25					
Basic	17-51	09-48				
Identify Tgts.	09-21	12-15	10-42			
Survival	15-48	13-42	44-71	07-29		
Technical	21-56	12-65	47-63	10-32	46-76	
FUNCTIONAL CATEGORIES						
Communications	100	21-28	17-51	09-21	15-50	21-58
Vehicle Ops.	21-28	100	09-48	12-15	22-28	20-35
Navigate	12-45	06-30	65-79	12-32	25-57	31-48
Field Tech.	09-46	04-27	36-93	08-39	13-63	24-55
Weapons	12-41	10-39	67-35	04-35	37-62	34-59
Anti Air/Tank Wpns.	14	-	32	20	26	-
Customs & Laws	13-33	11-30	56-67	03-20	31-47	36-44
Identify Tgts.	09-21	12-15	10-42	100	07-32	11-33
First Aid	09-35	12-25	31-55	06-26	63-98	30-73
NBC	15-51	11-41	41-62	05-26	78-89	39-61
Technical: 13B	18-21	-	47-56	18-24	42-51	75-97
19E	36	-	52-55	28-29	47-48	80-88
31C	34-49	14-35	32-57	13-29	38-51	65-81
63B	-	35-62	37-56	-	29-44	62-91
64C	-	-	55	11	50	100
71L	-	-	29-43	-	26-39	53-88
91A	-	01-13	20-55	-03-19	42-76	45-98
95B	20-31	06-20	33-53	12-17	28-46	63-85

Note: The numbers shown are the range of correlations that resulted for individual MOS; under the Technical functional category, however, the range of correlations is shown across the individual MOS Technical functional categories. Decimals have been omitted in the correlations.

Table 4-14

Correlations Between Criterion Factor Scores and Functional Categories for Hands-On Component

FACTORS	Commo.	Vehicle	Basic	Survival	Technical
Communications					
Vehicles	11-29				
Basic	06-26	07-15			
Survival	04-22	04-16	00-04		
Technical	06-28	07-15	12-42	10-29	

FUNCTIONAL CATEGORIES

	Commo.	Vehicle	Basic	Survival	Technical
Communications	100	10-29	05-26	02-20	07-30
Vehicle Ops.	10-29	100	07-15	11-16	08-11
Navigate	04-21	05-13	53-100	09-35	09-24
Field Tech.	08-18	05	39-70	08-13	09-18
Weapons	-01-22	-01-14	30-85	-01-31	07-37
Anti Air/Tank Wpns.	06	-	51	˙12	-
Customs & Laws	-	05	46	02	-02
First Aid	06-17	04-13	05-40	67-100	04-30
NBC	-04-17	02-12	06-22	46-81	04-22

		Commo.	Vehicle	Basic	Survival	Technical
Technical:	13B	08-09	-	26-42	12-16	66-95
	19E	18-21	-	16-19	16-23	80-82
	31C	13-31	04-11	12-26	00-18	55-76
	63B	-	07-13	06-07	01-05	47-82
	64C	-	-	12	11	100
	71L	-	-	10-20	10-11	44-93
	91A	-	-	01-23	00-32	39-96
	95B	07	08	17	12	100

Note: The numbers shown are the range of correlations that resulted for individual MOS; under the Technical functional category, however, the range of correlations is shown across the individual MOS Technical functional categories. Decimals have been omitted in the correlations.

Training Test Scores

Criterion scores for the training knowledge tests were derived in the same way as for the job knowledge tests. The results of the expert judgments and the exploratory factor analyses suggested that the six-score solution was also a reasonable one. Consequently, in the subsequent analyses aimed at developing a comprehensive model of job performance, the six content categories were scored in each of the three tests (hands-on, job knowledge, school knowledge) in each MOS in Batch A.

Basic Scores From the Rating Scales

For each soldier ratee in the sample, the goal was to obtain ratings from two supervisors and four peers who had worked with the ratee for at least two months and/or were sufficiently familiar with the ratee's job performance. The specific procedures used to identify peer and supervisor ratees can be found in Pulakos and Borman (1986). Overall, there were an average of 3.1 peer and 1.9 supervisor ratings for each ratee. The number of raters per ratee was sufficient to allow reasonable estimates of interrater reliability.

Raters did not succumb to excessive central tendency or leniency. The mean ratings were between 4 and 5 on the 7-point scales and the standard deviations were generally over 1.00.

Interrater Reliability

Interrater reliabilities were estimated with the intraclass correlation coefficient. In general, reliabilities of the individual scales were in the .30 - .45 range, and the reliabilities of the sums of the Army-wide and MOS-specific respectively were .65 and .55 using supervisor ratings. For peer ratings, the mean reliabilities were .58 and .42.

Factor Analysis of the Rating Scales

The reduction of the individual rating scales to a smaller set of aggregated scores was accomplished largely by means of exploratory factor analysis.

Army-Wide Performance Rating Scales. Principal factor analyses with a varimax rotation for the Army-wide scales were performed across MOS for peer raters, for supervisor raters, and for the combined peer and supervisor rater groups. Virtually identical results were obtained for all three rater groups, and a three-factor solution was chosen as the most meaningful. The names of the factors and the rating dimensions loading highest on each factor are shown in Table 4-15. Loadings for the rotated factor solutions and the combined group are shown in Table 4-16.

To determine how well the factor solution would hold up within individual MOS, factor scores using the factor scoring matrixes generated from the analyses across MOS were computed within the peer rater group, within the supervisor rater group, and for the combined peer and supervisor rater group.

124

Then, correlations were computed between the factor scores and the original behavioral dimension ratings. These analyses generally supported the stability and appropriateness of the three-factor structure across rating source and MOS.

Table 4-15

Army-Wide Performance Rating Scales Factors

Factor 1:	Job-Relevant Skills and Motivation
	Technical Knowledge/Skill
	Leadership
	Effort
	Self-Development
	Maintaining Equipment
Factor 2:	Personal Discipline
	Following Regulations
	Self-Control
	Integrity
Factor 3:	Physical Fitness and Military Bearing
	Military Appearance
	Physical Fitness

Table 4-16

Army-Wide Performance Rating Scales Three-Factor Solution for Combined Peer and Supervisor Raters

Rotated Factor Pattern*

Factor 1	Factor 2	Factor 3		Dimensions
.71	.28	.30	A:	Technical Skill
.69	.30	.37	E:	Leadership
.69	.43	.26	B:	Effort
.57	.38	.38	I:	Self-Development
.54	.34	.35	F:	Maintaining Equipment
.41	.69	.30	C:	Following Regulations
.22	.63	.20	J:	Self-Control
.50	.59	.28	D:	Integrity
.32	.32	.57	G:	Military Appearance
.21	.15	.49	H:	Physical Fitness

*Factor 1 - Job-Relevant Skills and Motivation; Factor 2 - Personal Discipline; Factor 3 - Physical Fitness and Military Bearing

MOS-Specific Performance Rating Scales. For the MOS-specific scales, principal factor analyses with a varimax rotation were conducted within MOS and separately for the peer and supervisor raters. The objective was to look for common themes that might be evident across MOS, even though different dimensions comprised each of the nine sets of scales.

Inspection of the factor analyses revealed a two-factor solution that could be used for all nine MOS. The rating dimensions loading highest on one of the factors consisted mainly of core job requirements and tasks, while those loading highest on the second factor were more peripheral job duties. Accordingly, for all MOS, a two-factor solution was chosen to represent the MOS-specific aspect of the criterion domain, with the factors named as follows: Core Responsibilities, and Other Responsibilities.

Combat Effectiveness Ratings. The combat scales were Army-wide summated scales based on the 40 items that survived the field tests and were designed to evaluate performance under degraded conditions and the increased confusion, workload, and uncertainty of a combat environment. A factor analysis of these items based on the combined samples from the Concurrent Validation suggested that two factors could be extracted. The first factor contained items that seemed to reflect performance under adverse, difficult, or dangerous conditions. The second was composed largely of items dealing with making mistakes, getting into trouble, or creating discipline problems. Consequently, items within each factor were summed to produce two scores for expected combat effectiveness: Performing Under Adverse Conditions and Avoiding Mistakes.

Army-Wide Common Task Ratings. The distributional properties, reliabilities, and factor structure of the 11 common task rating scales were analyzed using the same procedure as for the Army-wide performance scales. In general, these scales showed greater central tendency, lower reliabilities, and a less clear factor structure. Consequently, they were not used in the final criterion scoring.

Summary

To summarize the results of the rating scale score analyses:

- A three-factor solution (Job-Relevant Skills and Motivation, Personal Discipline, and Physical Fitness and Military Bearing) was chosen as the most psychologically meaningful for the Army-wide performance rating scales.

- Factor analyses of the MOS-specific rating scales yielded a two-factor solution across all nine MOS (Core Responsibilities, and Other Responsibilities).

- Factor analysis of the combat rating scales, using the combined sample, also produced a two-factor solution (Performing Under Adverse Conditions and Avoiding Mistakes).

MODELING OF CRITERION PERFORMANCE AND DEVELOPMENT OF
CRITERION FACTOR SCORES

Adding all the basic criterion scores into a single composite was viewed as too atheoretical, and developing a reliable and homogeneous measure of the general factor violated the basic notion that performance is multidimensional. A more formal way to model performance is to think in terms of its latent structure, postulate what that might be, and then resort to a confirmatory analysis.

Before any of the CV data were analyzed, the best speculation of the Project A staff had produced a preliminary model, shown in Figure 4-8. It went beyond what the Concurrent Validation data could examine and is included here only to illustrate the first stage in an almost continuous process of bootstrapping toward a more final conceptual description of the predictor/criterion space.

Successive revisions of the target model were then subjected to what might be described as "quasi" confirmatory analysis, using data from the Concurrent Validation sample. The purpose was to consider whether a single model of the latent structure of job performance would fit the data from all nine jobs. The analyses supporting this effort are summarized below.

Procedure

The results of the first level of aggregation have been referred to as the "basic" array of criterion scores. This reduced array of criterion variables is shown in Table 4-17. Because MOS do differ in their task content, not all 31 variables were scored in each MOS and there was some slight variation in the number of variables used in the subsequent analyses.

Table 4-17

Thirty-One Basic Criterion Scores Obtained by Aggregating Individual Rating Scales, Job Sample Tasks, Knowledge Test Items, and Archival Records

1.   Single scale rating of overall performance.

Three-Unit Weighted Factor Scores Obtained from the 10 Factor Analysis Army-Wide Behaviorally Anchored Rating Scales.

2.   Effort and leadership factor.
3.   Personal discipline factor.
4.   Physical fitness and military bearing factor.

Two-Unit Weighted Factor Scores Obtained Via Factor Analysis of the Job-Specific Behaviorally Anchored Rating Scales Developed for Each Job.

5.   Core responsibilities factor.
6.   Peripheral responsibilities factor.

(Continued)

Table 4-17 (Continued)

Thirty-One Basic Criterion Scores Obtained by Aggregating Individual Rating Scales, Job Sample Tasks, Knowledge Test Items, and Archival Records

---

Two-Unit Weighted Factor Scores Outlined from the Expected Combat Performance Summated Rating Scale.

7. Performing well under adverse conditions factor.
8. Avoiding mistakes factor.

Archival/Administrative Performance Indicators.

9. Awards and certificates.
10. Physical readiness test score.
11. M16 qualification score.
12. Articles 15/flag actions.
13. Promotion rate deviation score.

Task Proficiency Scale Scores Obtained by Clustering Items for Hands-On Job Sample Tests (HO).

14. Core technical (MOS-specific).
15. Communications.
16. Vehicle operation and maintenance.
17. General soldiering.
18. Identifying target and threat vehicles and aircraft.
19. Safety and survival.

Job Knowledge Scale Scores Obtained by Clustering Items From Job Knowledge Tests (JK).

20. Core technical (MOS-specific).
21. Communications.
22. Vehicle operation and main.
23. General soldiering.
24. Identifying target and threat vehicles and aircraft.
25. Safety and survival.

Training Knowledge Scale Scores Obtained by Clustering Items From Training School Knowledge Tests (SK).

26. Core technical (MOS-specific).
27. Communications.
28. Vehicle operation and maintenance.
29. General soldiering.
30. Identifying target and threat vehicles and aircraft.
31. Safety and survival.

128
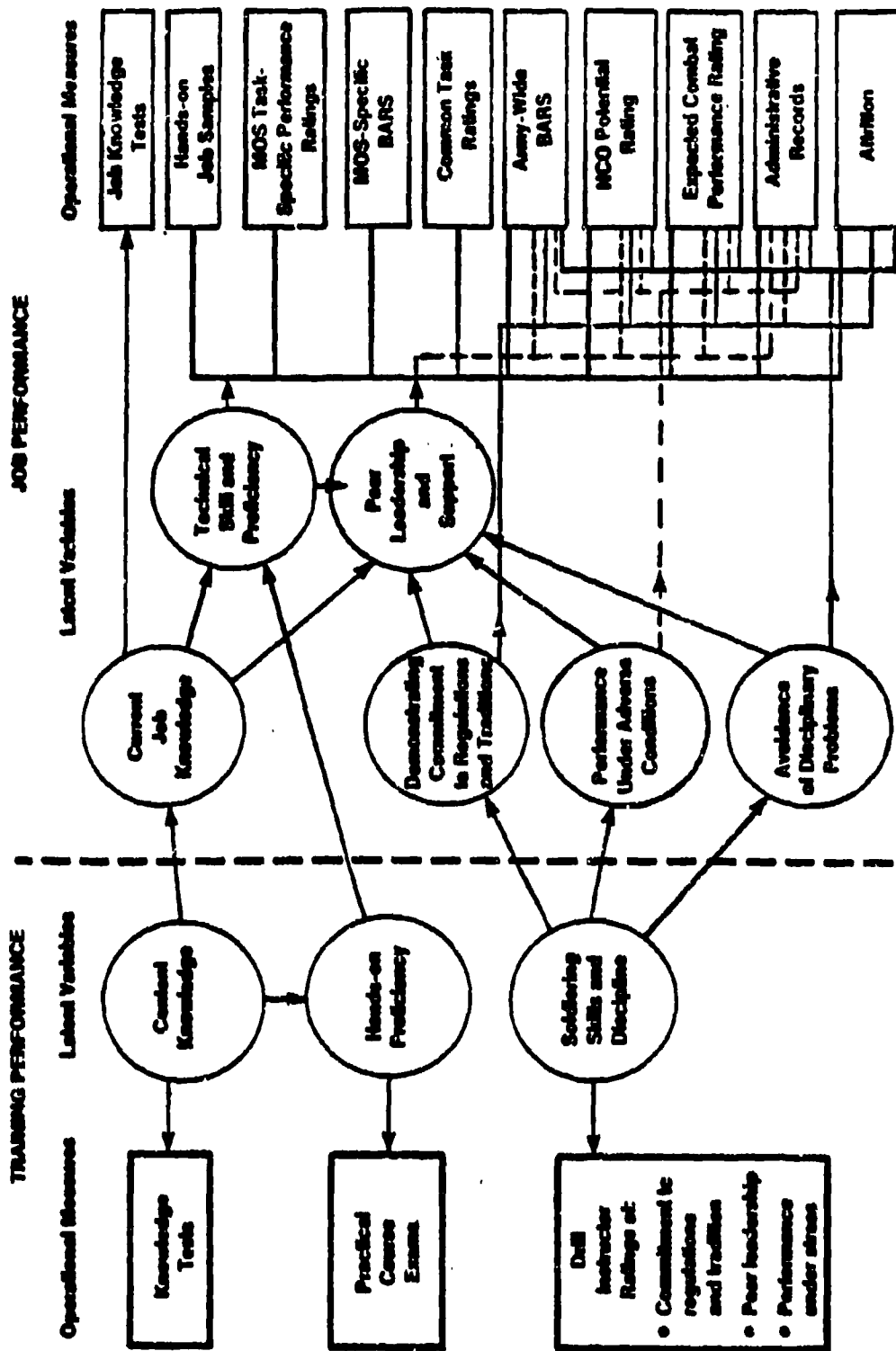
Figure 4-8. Preliminary model of enlisted job performance.

129

A Revised Model of Job Performance

Construction of a Target Model

The next step was to build a revised target model of job performance that could be tested for goodness-of-fit within each of the nine jobs, using the CV data. To do this, the intercorrelation matrixes of the basic criterion scores for the nine MOS were each subjected to another round of empirical factor analysis to suggest possible modifications.

Several consistent results were observed. First, as expected, there was the general prominence of "methods" factors, specifically one methods factor for the ratings and one methods factor for the written tests. Secondly, there was a close correspondence between the administrative measures scales and the three Army-wide rating factors. The awards and certificates scale from the administrative measures loaded together with the Army-wide Effort/Leadership rating factor; the Articles 15 score and the promotion rate scale loaded with the Personal Discipline factor.

Based on such findings, a revised model was constructed to account for the correlations among performance measures. It included five job performance constructs which are defined in Figure 4-9.

An issue that remained was whether the job-specific BARS were measuring job-specific technical knowledge and skill, or effort and leadership, or both. For purposes of model fitting the MOS-specific BARS core factor was hypothesized to load on both Core Technical and Effort/Leadership.

Another issue was whether it was necessary to posit hands-on and administrative measures "methods" factors to account for the inter-correlations within each of these sets of measures. Since the average intercorrelation among the scores within each of these sets was not particularly high, the hypothesized model did not include these two additional methods. However, it did include the ratings and written test methods factors. Consequently, the complete model specified the following seven factors:

1. Core Technical Proficiency
2. General Soldiering Proficiency
3. Effort and Leadership
4. Personal Discipline
5. Physical Fitness and Military Bearing
6. Ratings method factor
7. Paper-and-pencil method factor

130

1. Core Technical Proficiency

This performance construct represents the proficiency with which the soldier performs the tasks that are "central" to the MOS. The tasks represent the core of the job and they are the primary definers of the MOS. For example, the first-tour Armor Crewman starts and stops the tank engines; loads and unloads the main gun; boresights the M60A3; engages targets with the main gun; and performs misfire procedures. This performance construct does not include the individual's willingness to perform the task or the degree to which the individual can coordinate efforts with others. It refers to how well the individual can execute the core technical tasks the job requires, given a willingness to do so.

2. General Soldiering Proficiency

In addition to the core technical content specific to an MOS, individuals in every MOS also are responsible for being able to perform a variety of general soldiering tasks--for example, determines grid coordinates on military maps; puts on, wears, and removes M17 protective mask with hood; determines a magnetic azimuth using a compass; and recognizes and identifies friendly and threat aircraft. Performance on this construct represents overall proficiency on these general soldiering tasks. Again, it refers to how well the individual can execute general soldiering tasks, given a willingness to do so.

3. Effort and Leadership

This performance construct reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient: While appropriate knowledges and skills are necessary for successful performance, this construct is meant only to reflect the individual's willingness to do the job required and to be cooperative and supportive with other soldiers.

4. Personal Discipline

This performance construct reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this construct show a commitment to high standards of personal conduct.

5. Physical Fitness and Military Bearing

This performance construct represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

Figure 4-9. Definitions of the Job Performance Constructs.

131

Confirmation of the Model Within Each Job

The next step in the analysis was to conduct separate tests of goodness-of-fit of this target model within each of the nine jobs. This was done using the LISREL confirmatory factor analysis program (Joreskog & Sorbom, 1981).

As is not uncommon when using confirmatory models, some problems were encountered in fitting the hypothesized model to several of the jobs. Some factor loadings were greater than one, with negative uniqueness estimates for the corresponding observed variables. Also, estimates of the correlations among the performance constructs occasionally exceeded unity. These problems necessitated a certain amount of ad hoc cutting and fitting in the form of computing the squared multiple correlation (SMC) for predicting each observed variable from all of the other variables, and setting the uniqueness estimates (i.e., Theta-Epsilon diagonal) to 1.0 minus this SMC. This approach eliminated all factor loadings and correlations greater than one. In most cases, a second "iteration" was performed to adjust the initial uniqueness estimates (Theta-Epsilon) so that the diagonal of the estimated correlation matrix would be as close to 1.0 as possible. The final factor loading estimates for each job are shown in Table 4-18.

LISREL also computes a goodness-of-fit index based on a comparison of the actual correlations among the observed variables and the estimated correlations. The goodness-of-fit is distributed as chi-square, with degrees of freedom dependent on the number of observed variables and the number of parameters estimated. The expected value of chi-square is equal to the degrees of freedom; it is a sign that the model does not fit the correlations among the observed variables.

However, the chi-square values should be interpreted with caution because the hypothesized target model was based in part on analyses of these same data. In addition, LISREL was "told" that the Theta-Epsilon (uniqueness) parameters all were fixed, and therefore did not "use up" degrees of freedom estimating these parameters; in fact, these values were estimated entirely from the data.

Confirmation of an Overall Model

The results of the confirmatory procedures applied to the performance measures from each job generally supported a common structure of job performance. A final step was to determine whether the variation in some of these parameters across jobs could be attributed to sampling variation by hypothesizing that (a) the correlation among factors was invariant across jobs, and (b) the loadings of all of the Army-wide measures on the performance constructs and on the rating method factor were also constant across jobs.

132

Table 4-18

Factor Loadings: Separate Model of Job Performance for Each Job

Construct/Factor[a]	MOS								
	11B	13B	19F	31C	63B	64C	71L	91A	95B
Core Technical									
HO Technical	--	.61	.47	.64	.51	.29	.77	.59	.32
JK Technical	--	.75	.78	.79	.74	.26	.78	.75	.32
SK Technical	--	.70	.79	.73	.82	.55	.22	.81	.43
MOS Tech Rating	--	.45	.10	.22	.25	.25	.34	.10	.13
General Soldiering									
HO Soldier	.60	.51	.46	.64	.17	.50	.60	.42	.60
HO Safety	.26	.33	.32	.31	.12	.63	.37	.48	.47
HO Communications	.05	.06	.39	.56	--	--	--	--	.80
HO Vehicle	--	--	--	.22	.17	b	--	--	.31
JK Soldier	.76	.52	.74	.62	.45	.48	.87	.58	.46
JK Safety	.55	.37	.75	.38	.71	.51	.72	.58	.33
JK Communications	.30	.23	.65	.38	--	--	--	--	.29
JK Vehicle	--	.17	--	.10	.41	b	--	--	.35
JK Identify	.46	--	.20	.28	--	.12	--	.24	.21
SK Soldier	.73	.45	.67	.39	.78	.56	.45	.44	.42
SK Safety	.47	.32	.53	.62	.57	.47	.30	.64	.32
SK Communications	.42	.26	.42	--	.41	.35	.20	--	.20
SK Vehicle	.22	.24	.05	.30	.61	b	.22	.47	.28
SK Identify	.46	--	.46	.13	--	--	--	--	--
Effort/Leadership									
Eff/Ldr Rating	.76	.56	.85	.64	.68	.83	.66	.76	.70
MOS Tech Ratings	.70	--	.63	.40	.41	.50	.25	.59	.52
MOS Other Rating	.77	.41	.48	.43	.54	.62	.43	.61	.56
Combat Exmplry	.80	.47	.68	.54	.57	.87	.63	.80	.77
Combat Problems	.48	.20	--	.39	.52	.53	.55	--	.56
Awards/Certificate	.32	.23	.24	.19	.28	.25	.34	.04	.22
Overall Rating	.46	.39	.33	.17	.57	.42	.65	--	.41
Discipline									
Discipline Rating	.77	.58	.73	.45	.63	.85	.74	.58	.73
Combat Problems	.23	.16	.62	.03	.05	.19	--	.02	.33
Articles 15	-.63	-.61	-.55	-.62	-.65	-.47	-.69	-.46	-.50
Promotion Rate	.74	.61	.68	.79	.63	.57	.59	.54	.54
Overall Rating	.39	.20	.53	.54	.09	.42	.06	.75	.38

(Continued)

133

Table 4-18 (Continued)

Factor Loadings: Separate Model of Job Performance for Each Job

Construct/Factor[a]	MOS								
	11B	13B	19E	31C	63B	64C	71L	91A	95B
Fitness/Bearing									
Fitness Rating	.69	.23	.84	.48	.54	.42	.50	.60	.78
Physical Readiness	.11	.90	.49	.89	.70	.53	.76	.09	.69
Ratings Method									
AW Ratings	.60	.73	.47	.70	.66	.54	.65	.66	.66
MOS Ratings	.73	.73	.60	.69	.67	.49	.69	.54	.63
Combat Ratings	.47	.65	.55	.69	.57	.27	.55	.47	.40
Written Method									
JK Technical	--	.47	.28	.55	.59	.73	.44	.58	.57
JK Soldier	.41	.51	.33	.40	.61	.57	.11	.37	.59
JK Safety	.37	.52	.12	.63	.08	.49	.17	.76	.57
JK Communications	.34	.11	.07	.55	--	--	--	--	.52
JK Vehicle	--	--	--	.42	.62	[b]	--	.24	.21
JK Identify	-.15	.23	.50	.36	--	.05	--	.08	.23
SK Technical	--	.48	.48	.55	.46	.88	.42	.27	.50
SK Soldier	.50	.66	.54	.59	.15	.51	.54	--	.54
SK Safety	.53	.55	.42	.29	.34	.48	.44	.19	.60
SK Communications	.51	.47	.46	--	.16	.24	.05	--	.42
SK Vehicle	.48	.55	[b]	.38	.05	.42	.49	.57	.24
SK Identify	.21	--	.42	.44	--	--	--	--	--
M16 Qualification	.71	.71	.71	.71	.71	.71	.71	.71	.71

[a]HO = Hands-on; JK = Job Knowledge; SK = School Knowledge; AW = Army-Wide.

[b]Vehicle content was merged into the Core Technical factor for MOS 64C.

The proposed overall model was a relatively stringent test of a common latent structure since it was quite possible that selectivity differences in the different jobs would tend to make it appear that the different jobs require different performance models, when in fact they do not. However, the over-all model fit very well. The root mean square residual was .047, and chi-square was 2508.1 with 2403 degrees of freedom after adjusting for missing variables and the use of the data in estimating uniqueness. Table 4-19 shows the final mapping of the criterion measures on the five performance components.

Obtaining Criterion Factors Scores for Individuals

To obtain an individual's score on each of the five constructs, the variables composing each factor were scored and combined in the following manner.

The Core Technical Proficiency construct is operationally defined as the standardized sum of the MOS-specific technical task content from the hands-on tests, the job knowledge tests, and the school knowledge tests.

The General Soldiering Proficiency score is also composed of two major components, each of which is standardized and then added to generate the criterion score. The first component is operationally defined as the sum of the CVBIS[2] scores from the hands-on test, and the second component is defined as the sum of the CVBIS scores from both the job knowledge and school knowledge tests.

The Effort/Leadership criterion factor is composed of four major components, each of which is standardized before the four are summed. The first component corresponds to the single rating for Overall Effectiveness. The second component is composed of three subcomponents. The first is one of the three factor scores derived from the Army-wide BARS scales (i.e., the Army-wide Effort/Leadership factor) and consists of the unit-weighted sum of five different scales (Technical Skill; Effort; Leadership; Maintain Equipment; Self Development). The second and third subcomponents are the two factor scores derived from the MOS-specific BARS rating scales. (It should be noted that all rating scores used in the computation of all criterion constructs are the average of the ratings provided by supervisors and peers.) The third component is the average of the two combat rating scales. Finally, the fourth component corresponds to the administrative measure identified as Total Awards/Letters.

---

[2] A set of content categories derived from the hands-on and knowledge test variables, where tasks and items were assigned as follows: Communication (radio operation); Vehicle Maintenance; Basic Soldiering Skills (field techniques, weapons, navigation, customs and law); Identify (friendly and enemy aircraft and vehicles); Technical Skills (specific to the job); Safety/Survival (first aid, NBC).

Table 4-19

Mapping of Performance Factors Onto Latent Performance Constructs

| | Latent Performance Constructs | | | | | | | |
| | Content Constructs | | | | | Method Constructs | | |
Criterion Name[a]	Core Technical Proficiency	General Soldiering Proficiency	Effort/ Leadership	Personal Discipline	Physical Fitness/ Military Bearing	Written Know- ledge Test	Rating Scale	M16 Qualification
AF Effort			X					
AFa Discipline				X			X	
AFb Fitness			X	X	X		X	
AFc Overall							X	
AM Technical							X	
AMa Other			X	X			X	
CAE Perform Well			X	X			X	
CAbc Avoid Mistake			X				X	
AD Awards/Cert			X		X			
ADa Phys Readiness					X			
ADb M16								X
ADc Articles 15				X				
ADd Promotion Rate				X				
BD Tactical	X							
BD Communications		X						
BD Vehicles		X						
BD General Soldier		X						
BD ID Threat/Target		X						
BD Safety/Survival		X						
JK Tactical	X					X		
JK Communications		X				X		
JK Vehicles		X				X		
JK General Soldier		X				X		
JK ID Threat/Target		X				X		
JK Safety/Survival		X				X		
SK Technical	X					X		
SK Communications		X				X		
SK Vehicles		X				X		
SK General Soldier		X				X		
SK ID Threat/Target		X				X		
SK Safety/Survival		X				X		

Note: Within each rating instrument, all of the factors were constrained to have an equal loading on the Rating Scales method construct. For example, the Perform Well and Avoid Mistakes factors from the Combat Performance Prediction Scale were constrained to have identical loadings on the Rating Scales method construct, but this loading did not set here to be to be the same as the loading for the Army-Wide BARS factors, the BARS-Specific BARS factors, or the Common Task Scales factors.

[a]AW = Army-wide behaviorally anchored rating scales; BD = hands-on; JK = job knowledge; SK = school knowledge.

The Personal Discipline factor is composed of two major components, each of which is standardized before the two are added. The first component is the Personal Discipline score derived from Army-wide BARS and consists of the unit-weighted sum of three different scales (Following Regulations; Integrity; Self-Control). The second component is the sum of two administrative measures, Articles 15/Flag Actions and Promotion Rate Deviation score.

The fifth criterion factor, Physical Fitness and Military Bearing, is composed of two components; again, each is standardized before they are added to generate a criterion score. The first component is the Physical Fitness and Bearing score derived from the Army-wide BARS and consists of the unit-weighted sum of two different scales (Military Appearance; Physical Fitness). The second component corresponds to the administrative measure identified as the Physical Readiness score.

Five residual scores were then created from the five criterion factors by partialing the paper-and-pencil methods factor from Core Technical and General Soldiering and the ratings methods factor from Effort/Leadership, Personal Discipline, and Fitness and Bearing.

Criterion Intercorrelations

The five criterion factor scores, the five criterion residual scores, the single rating obtained from the overall performance rating scales, and the total score from the hands-on tests were used to generate a 12 x 12 matrix of criterion intercorrelations for each MOS in Batch A. The averages of these correlations across MOS are shown in Table 4-20. The inter- correlations between factor scores within method (factor 1 with 2 or 3 with 4) are higher, as expected, than factor pairs which do not confound method (e.g., 1 with 3 or 2 with 4). However, they are not so high that collapsing the five factors into some smaller number would be justified. In fact, factors 1 and 2, which intercorrelate .53 on the average, yield different profiles of correlations with the tests in the predictor battery.

Assuming a reliability of about .60 for each measure would yield an intercorrelation of about .34 for the correlation of the overall performance rating with the total hands-on score when corrected for attenuation. A reasonable conclusion is that while performance on a standardized job sample is a significant component of performance, it is by no means all of it.

The correlations of the residualized factor 3 (Effort/Leadership residual) with the Core Technical factor, the Core Technical residual, the General Soldiering Proficiency factor, the overall rating scale, and the hands-on total score all are about the same. Also, as compared to the correlation of the Effort/Leadership raw scores with these same variables, the correlations of the Effort/Leadership residual with the Core Technical and General Soldiering Proficiency factors go up while the correlations with Personal Discipline and Physical Fitness go down. Residualizing factor 3 (by removing the ratings method factor) makes it more like a "can do" factor and less like a "will do" factor.

137

Table 4-20

Mean Intercorrelations Among 12 Summary Criterion Measures for the Batch A NOS

Criterion Summary Score	CTP Raw	GSP Raw	E/L Raw	PD Raw	F/B Raw	OPF	HOT	CTP Res	GSP Res	E/L Res	PD Res	F/B Res
Core Tech Prof (raw)	1.00	.53	.28	.19	.03	.24	.74	.88	.38	.47	.23	.04
Gen Soldier Prof (raw)	.53	1.00	.27	.16	.04	.21	.72	.39	.89	.45	.19	.05
Effort/Leadership (raw)	.28	.27	1.00	.59	.46	.87	.26	.35	.33	.65	.28	.19
Personal Discipline (raw)	.19	.16	.59	1.00	.33	.65	.15	.26	.23	.44	.89	.19
Fitness/Bearing (raw)	.03	.04	.46	.33	1.00	.47	.07	.03	.04	.25	.17	.92
Overall Perf Rating	.24	.21	.87	.65	.47	1.00	.20	.31	.26	.44	.33	.19
Hands-On Total	.74	.72	.26	.15	.07	.20	1.00	.82	.79	.44	.18	.09
Core Tech Prof (resid)	.88	.39	.35	.26	.03	.31	.82	1.00	.44	.45	.25	-.01
Gen Soldier Prof (resid)	.38	.89	.33	.23	.04	.26	.79	.44	1.00	.43	.21	.01
Effort/Leadership (resid)	.47	.45	.65	.44	.25	.44	.44	.45	.43	1.00	.48	.28
Personal Discipline (resid)	.23	.19	.28	.89	.17	.33	.18	.25	.21	.48	1.00	.20
Fitness/Bearing (resid)	.04	.05	.19	.19	.92	.19	.09	-.01	.01	.28	.20	1.00

Concluding Comments

In general, these intercorrelations seem to behave in very lawful ways and are consistent with a multidimensional model of performance. In spite of some confounding of factor content with measurement method, the latent performance structure appears to be composed of very distinct components and it is reasonable to expect that the different performance constructs would be predicted by different things. Since (a) the five-factor solution is stable across jobs sampled from this population, (b) the performance constructs seem to make sense, and (c) the constructs are based on measures carefully developed to be content valid, it seemed safe to ascribe some degree of construct validity to them.

BASIC CONCURRENT VALIDATION RESULTS

As described previously, 24 scores were used to assess the predictor domain and five criterion construct scores were developed to provide a comprehensive assessment of job performance. Consequently, the basic validation data generated by the Concurrent Validation are contained in the 24 x 5 correlation matrix that could be computed for each MOS in the sample.

The predictor scores were grouped into six domains and the multiple correlation of the predictor scores within each domain with each of the criterion construct scores was computed for each of the nine MOS in Batch A. Figure 4-10 depicts the relationships that were expected between the predictor domains and the five job performance constructs. Each R was corrected for range restriction using the multivariate procedure described in Lord and Novick (1968) and adjusted for shrinkage using the procedure described by Claudy (1978).

Initial Multiple Correlation Results

Given six predictor domains and five job performance constructs, 30 multiple correlations were generated for each MOS. The mean validity (R) values for the nine MOS are reported in Table 4-21.

As a test of the hypothesized predictor-criterion relationships presented in Figure 4-10, the predictor composites were grouped into the two prescribed sets. For each set the R was computed with each of the five job performance constructs within each of the nine jobs. Mean Rs from these analyses are presented in Table 4-22. The pattern of correlations is very similar to that predicted in Figure 4-10. The one surprising result is the high correlation between the non-cognitive predictors and the two "can do" performance constructs. This is due primarily to the validity of the AVOICE, which has important implications for the development of optimal classification algorithms.

139

PREDICTOR DOMAIN   **CRITERION CONSTRUCT**

Cognitive Portion

| General Cognitive Ability |
| Spatial Ability |
| Perceptual-Psychomotor Ability |

| Core Technical Proficiency |
| General Soldiering Proficiency |

| Effort and Leadership |

Non-Cognitive Portion

| Temperament |
| Vocational Interests |
| Job Reward Preferences |

| Personal Discipline |
| Physical Fitness and Military Bearing |

Figure 4-10. Hypothesized predictor-criterion relationships.

Table 4-21

Mean Validity[a] for the Composite Scores Within Each Predictor Domain
Across Nine Army Enlisted Jobs

Job Performance Construct	Predictor Domain					
	General Cognitive Ability (K=4)[b]	Spatial Ability (K=1)	Perceptual-Psychomotor Ability (K=6)	Temper-ament (K=4)	Vocational Interests (K=6)	Job Reward Prefer (K=3)
Core Technical Proficiency	.63	.56	.53	.25	.35	.29
General Soldiering Proficiency	.65	.63	.57	.25	.34	.30
Effort and Leadership	.31	.25	.26	.33	.24	.19
Personal Discipline	.16	.12	.12	.32	.13	.11
Physical Fitness and Military Bearing	.20	.10	.11	.37	.12	.11

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]K is the number of predictor scores.

Table 4-22

Mean Validity[a] for the Cognitive, Non-Cognitive, and All Predictor
Composites Across Nine Army Enlisted Jobs

Job Performance Construct	Predictor Domain		
	Cognitive (K=11)[b]	Non-Cognitive (K=13)	All (K=24)
Core Technical Proficiency	.65	.44	.67
General Soldiering Proficiency	.69	.44	.70
Effort and Leadership	.32	.38	.44
Personal Discipline	.17	.35	.37
Physical Fitness and Military Bearing	.23	.38	.42

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]K is the number of predictor scores.

Incremental Validity

An important question is how to improve upon the validity of decisions made using the current selection and classification instrument. The validity of the General Cognitive Ability scores (computed from the ASVAB) was compared to the validity obtained when the scores from other predictor domains were added. The resulting mean validities are reported in Table 4-23.

Table 4-23

Mean Incremental Validity[a,b] for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs

	Predictor Domain					
Job Performance Construct	General Cognitive Ability (K=4)[c]	General Cognitive Ability Plus Spatial Ability (K=5)	General Cognitive Ability Plus Perceptual Psychomotor Ability (K=10)	General Cognitive Ability Plus Temperament (K=8)	General Cognitive Ability Plus Vocational Interests (K=10)	General Cognitive Ability Plus Job Reward Pref (K=7)
Core Technical Proficiency	.63	.65	.64	.63	.64	.63
General Soldiering Proficiency	.65	.68	.67	.66	.66	.66
Effort and Leadership	.31	.32	.32	.42	.35	.33
Personal Discipline	.16	.17	.17	.35	.19	.19
Physical Fitness and Military Bearing	.20	.22	.22	.41	.24	.22

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]Incremental validity refers to the increase in R afforded by the new predictors above and beyond the R for the Army's current predictor battery, the ASVAB.
[c]K is the number of predictor scores.

None of the predictor domains added more than .03 to the prediction of Core Technical Proficiency or General Soldiering Proficiency. In both instances, the composite that added the incremental validity was Spatial Ability. However, the four Temperament predictor scores added .11 to the predicting of Effort and Leadership, .19 to Personal Discipline, and .21 to Physical Fitness and Military Bearing.

Overall, the results are consistent with the hypotheses that: (a) cognitive ability would be the most valid predictor of Core Technical Proficiency and General Soldiering Proficiency; (b) non-cognitive composites would be the most valid predictors of Personal Discipline and Physical Fitness and Military Bearing; and (c) both cognitive and non-cognitive predictors would be useful for predicting Effort and Leadership.

Predictor Relationships With Criterion Residual Scores

Another method of studying the construct validity of both predictors and criteria is to examine how the pattern of predictor-criterion relationships changes when the variance attributable to the methods factors is removed from the five performance construct scores. These results are presented in Table 4-24.

To compute residual performance construct scores, the variance attributable to the written test factor was partialed from the scores for Core Technical Proficiency and General Soldiering Proficiency, and the variance attributable to the rating factor was partialed from the scores for Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing.

The table shows that the residual scores for Core Technical Proficiency and General Soldiering Proficiency were less predictable than the raw scores. However, the level of prediction is still substantial even when all variance attributable to the paper-and-pencil measurement mode is partialed out. One strong conclusion is that measurement method does not explain away the validity of ASVAB.

For Effort and Leadership, the cognitive predictor scores predicted the residual performance construct scores better than they predicted the raw performance construct scores. For example, the mean R of the General Cognitive Ability composite rose from .31 to .46. The increase was .16 for Spatial composite and .12 for the Perceptual-Psychomotor composite. For the ABLE composite, the results were reversed and the multiple correlation decreased from .33 to .31. The Vocational Interests composite and the Job Reward Preferences composite "behaved" similarly to the Cognitive Ability composite. The mean Rs were greater for the residual Effort and Leadership score than for the raw Effort and Leadership score.

143

Table 4-24

Mean Validity[a] for the Composite Scores Within Each Predictor Domain
Across Nine Army Enlisted Jobs

Job Performance Construct	Type of Score	Predictor Domain					
		General Cognitive Ability (K=4)[b]	Spatial Ability (K=1)	Perceptual-Psychomotor Ability (K=6)	Temper-ament (K=4)	Voc Inter (K=6)	Job Reward Pref (K=1)
Core Technical Proficiency	Raw	.63	.56	.53	.26	.35	.29
	Resid	.47	.37	.37	.22	.28	.21
General Soldiering Proficiency	Raw	.65	.63	.57	.25	.34	.30
	Resid	.49	.48	.41	.21	.26	.22
Effort and Leadership	Raw	.31	.25	.26	.33	.24	.19
	Resid	.46	.41	.38	.31	.32	.27
Personal Discipline	Raw	.16	.12	.12	.32	.13	.11
	Resid	.19	.15	.13	.28	.15	.10
Physical Fitness and Military Bearing	Raw	.20	.10	.11	.37	.12	.11
	Resid	.21	.11	.14	.35	.14	.10

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]K is the number of predictor scores.

144

This pattern of correlations for Effort and Leadership suggests two interesting conclusions. First, it provides additional evidence that the Vocational Interests scores are more similar to cognitive predictors than to temperament predictors. Second, the changes in correlations suggest that Effort and Leadership becomes more like a "can do" performance construct when the rating method factor is partialed out. However, the residual Effort and Leadership score continues to reflect the "will do" portion of the job performance space as suggested by its highest Rs. Thus, the residual Effort and Leadership score appears to tap both "can do" or maximal job performance and "will do" or typical job performance.

Partialing the rating factor from the Personal Discipline and the Physical Fitness and Military Bearing scores had little impact on the correlations of these scores with the predictor composites.

Stepwise multiple regression solutions within each of the six categories of predictor constructs are shown in Tables 4-25 and 4-26. The regression equations in Table 4-25 were computed on the combined samples from the nine MOS in Batch A for each of the last four Army-wide performance factors (i.e., General Soldiering, Effort/Leadership, Personal Discipline, and Physical Fitness/Military Bearing). The coefficients were computed on the combined samples because a series of analyses of variance had shown few Predictor by MOS interactions when the dependent variable was one of the four Army-wide factors. However, the profile of regression coefficients for predicting the Core Technical Proficiency factor was significantly different across MOS. The MOS by MOS stepwise regression solutions within predictor category are shown in Table 4-26.

For the four Army-wise components, some comparisons of interest are the following:

- Among ASVAB scores the quantitative and technical scores contribute the most to the prediction of General Soldiering Proficiency. The verbal score plays a more prominent role in the prediction of the Core Technical performance factor.

- While ASVAB does not contribute much to the prediction of performance factors 4 and 5, the ASVAB technical score does make a relatively large contribution to the prediction of factor 3, the Effort/Leadership factor.

- The differential contributions of the temperament (ABLE) scores to prediction of performance factors 3, 4, and 5 are clear, significant, and pronounced. The profiles look like they should.

- The combat interests score was the most predictive interest score among the scores generated from the AVOICE.

145

Table 4-25

Results of Stepwise Regressions Within Each Predictor Domain for the Four Army-Wide Performance Constructs Across All Nine Batch A MOS

Predictor Construct	Criterion Construct				
	General Soldiering (raw score)	Effort and Leadership (resid score)	Effort and Leadership (raw score)	Personal Discipline (raw score)	Phys Fitness Mil Bearing (raw score)
ASVAB Factors					
Verbal	.10	.03	-.07	-.03	-.11
Quantitative	.20	.08	.03	.07	.03
Technical	.26	.21	.21	.06	-.05
Speed	.03	.07	.09	.04	.10
ADJ. UNCORR R	.461	.280	.206	.106	.161
Spatial					
Overall Spatial	.47	.25	.14	.07	-.05
UNCORRECTED R	.466	.253	.142	.068	.047
Computer					
Complex Perc Speed	-.09	-.06	-.07	--	--
Complex Perc Accy	.19	.07	.09	.05	--
Number Speed/Accy	-.14	-.06	-.09	-.03	--
Psychomotor	-.19	-.08	-.10	--	--
Simp Reaction Accy	.04	--	--	--	-.06
Simp Reaction Speed	--	--	--	--	-.07
ADJ. UNCORR R	.363	.149	.208	.032	.071
Temperament					
Adjustment	.09	.04	.03	.03	--
Dependability	.04	--	.06	.30	.12
Achievement	.04	.23	.25	--	.12
Phys Condition	-.06	--	--	-.06	.24
ADJ. UNCORR R	.129	.255	.303	.303	.356
Interests					
Combat	.24	.20	.17	--	.04
Machines	--	--	--	-.04	-.06
Audiovisual	--	--	-.04	--	--
Technical	--	.06	.08	.09	.14
Food Service	-.10	-.16	-.12	.06	-.05
Protective Svc	-.06	--	--	-.09	--
ADJ. UNCORR R	.229	.235	.199	.078	.119
Job Values					
Security	--	.03	.05	.05	.10
Autonomy	.05	.07	.03	-.06	-.05
Routine	-.11	-.12	-.09	-.03	-.02
ADJ. UNCORR R	.123	.150	.112	.063	.097

Table 4-26

Results of Stepwise Regressions Within Each Predictor Domain for MOS-Specific Core Technical Proficiency for Each of the Nine Batch A MOS

	MOS								
Predictor Construct	11B	13B	19E	31C	63E	64C	71L	91A	95B
ASVAB Factors									
Verbal	.20	--	.13	.19	--	--	.16	.25	.11
Quantitative	.14	.09	.15	.14	--	.14	.38	.12	.16
Technical	.23	.23	.27	.23	.55	.34	.11	.19	.11
Speed	.10	--	--	.11	--	--	.08	.17	.09
ADJ, UNCORR R	.503	.254	.452	.427	.538	.413	.441	.456	.282
Spatial									
Overall Spatial	.48	.33	.43	.32	.41	.37	.41	.38	.28
UNCORRECTED R	.475	.334	.432	.315	.412	.366	.411	.380	.275
Computer									
Complex Perc Speed	-.25	-.10	--	--	-.08	-.14	--	--	--
Complex Perc Accy	.29	.11	.16	.13	--	.19	.27	.09	.13
Number Speed/Accy	-.11	-.11	-.20	-.25	-.08	-.07	-.22	-.20	-.19
Psychomotor	-.13	-.17	-.11	-.09	-.20	-.10	--	-.15	-.09
Simp Reaction Accy	--	--	.12	--	.08	.07	--	.08	--
Simp Reaction Speed	--	--	--	--	--	--	--	--	--
ADJ, UNCORR R	.406	.257	.343	.253	.242	.269	.325	.261	.228
Temperament									
Adjustment	--	.12	.14	--	.10	--	--	.10	.08
Dependability	--	--	.08	.10		--	.10	.19	.12
Achievement	.19	--	--	--	.09	--	.14	--	--
Phys Condition	--	--	-.13	--	-.12	--	-.10	-.15	--
ADJ UNCORR R	.143	.000	.129	.000	.119	.000	.176	.211	.114
Interests									
Combat	.25	.25	.26	--	.11	.09	.12	.18	--
Machines	--	.10	--	.13	.38	.09	-.23	--	--
Audiovisual	--	--	--	--	-.11	--	--	--	-.08
Technical	.08	--	--	.10	--	--	.19	--	--
Food Service	-.22	-.16	-.11	—	-.10	-.12	-.07	--	-.06
Protective Svc	-.11	-.10	--	--	-.14	--	--	--	--
ADJ, UNCORR R	.276	.255	.218	.000	.441	.135	.160	.039	.000
Job Values									
Security	--	--	--	--	--	--	--	.14	--
Autonomy	.08	.17	--	--	.14	.11	--	--	--
Routine	-.15	-.14	-.21	--	-.10	-.07	-.12	--	-.08
ADJ, UNCORR R	.141	.201	.166	.000	.133	.080	.038	.058	.000

For the MOS by MOS stepwise regression coefficient profiles used to predict the Core Technical factor (i.e., Table 4-26), the greatest differential is within the ASVAB and the AVOICE, and to a lesser extent within the spatial and computerized tests.

To look at the coefficients in another way, stepwise regressions were carried out when all 24 predictor scores were used to predict each performance factor. Again, the analyses for the four Army-wide criterion factors were carried out on a combined sample while the analyses against the Core Technical factor were done MOS by MOS. The results are shown in Tables 4-27 and 4-28.

Again the differential patterns appear across the four Army-wide performance factors and across MOS for the Core Technical factor. However, a surprise was the strong role played by the spatial and the combat interest constructs in predicting the technical performance factor in the combat specialties.

To round out the picture, the zero-order correlations (validity coefficients) corresponding to the regression coefficients in Tables 4-27 and 4-28 are shown in Tables 4-29 and 4-30.

Summary

At this point, Project A had reached a number of its basic goals.

- Multiple criterion measures had been developed and used to formulate five components of job performance.

- ASVAB was shown to be a highly valid predictor of job performance as reflected in the Core Technical performance and General Soldiering performance components.

- There was a considerable differential prediction for the total test battery across the five performance components within each MOS.

- The non-cognitive predictors added significantly to the prediction of the "will-do" components of performance and should prove to be valuable additions to the total system.

- As was expected, differential prediction across MOS was limited largely to the Core Technical performance factor. Both the ASVAB and the new experimental cognitive tests should contribute to differential prediction equations across major MOS clusters. However, the full analyses necessary to determine the prediction equations remain to be done.

Table 4-27

Results of Stepwise Regressions for the Four Army-Wide Performance Constructs Across All Nine Batch A MOS

Predictor Construct	Criterion Construct				
	General Soldiering (raw score)	Effort and Leadership (resid score)	Effort and Leadership (raw score)	Personal Discipline (raw score)	Phys Fitness Mil Bearing (raw score)
ASVAB Factors					
Verbal	.09	.03	-.06	--	-.10
Quantitative	.09	.04	--	.05	--
Technical	.12	.11	.15	.07	-.03
Speed	--	.04	.06	.03	.08
Spatial					
Overall Spatial	.25	.13	--	--	--
Computer					
Complex Perc Speed	--	--	-.05	--	--
Complex Perc Accy	.08	--	.04	--	--
Number Speed/Accy	-.02	--	--	.03	--
Psychomotor	-.04	--	-.02	--	--
Simp Reaction Accy	--	--	--	--	-.04
Simp Reaction Speed	-.03	--	--	--	-.05
Temperament					
Adjustment	--	--	--	--	--
Dependability	.11	.06	.11	.30	.09
Achievement	-.04	.15	.20	.03	.14
Phys Condition	--	.03	--	-.05	.22
Interests					
Combat	.13	.15	.10	--	.04
Machines	--	--	--	--	-.05
Audiovisual	--	-.02	-.04	-.03	.04
Technical	--	--	--	--	--
Food Service	-.04	-.08	-.06	-.04	--
Protective Svc	--	.03	--	-.03	-.05
Job Values					
Security	--	--	--	--	--
Autonomy	--	--	--	-.05	-.04
Routine	-.03	-.04	-.03	--	--
ADJ, UNCORR R	.540	.392	.366	.317	.385

Table 4-28

Results of Stepwise Regressions for MOS-Specific Core Technical Proficiency for Each of the Nine Batch A MOS

Predictor Construct	MOS								
	11B	13B	19E	31C	63B	64C	71L	91A	95B
ASVAB Factors									
Verbal	.17	--	.10	.21	--	--	.08	.26	.13
Quantitative	.09	--	--	.30	--	--	.27	--	--
Technical	.10	--	.16	--	.35	.30	-.13	.12	--
Speed	--	--	--	--	--	-.07	--	.13	--
Spatial									
Overall Spatial	.20	.25	.19	--	.14	.16	.25	.23	.22
Computer									
Complex Perc Speed	.18	--	--	--	--	-.12	--	--	--
Complex Perc Accy	.13	--	.09	-.10	--	.14	.15	--	.09
Number Speed/Accy	--	--	-.09	--	--	--	--	--	-.11
Psychomotor	--	--	--	--	--	--	--	--	--
Simp Reaction Accy	--	--	.07	--	--	--	--	--	--
Simp Reaction Speed	--	-.10	--	--	-.11	--	--	--	--
Temperament									
Adjustment	-.08	--	--	-.09	--	--	--	--	--
Dependability	.12	--	.10	.15	.13	.07	.11	.22	.12
Achievement	--	--	--	--	--	--	--	--	--
Phys Condition	--	--	-.09	--	-.06	--	--	-.13	--
Interests									
Combat	.15	.21	.17	--	--	--	-.16	.16	--
Machines	--	--	--	.21	.32	--	--	--	--
Audiovisual	--	--	--	--	-.14	--	--	-.09	-.13
Technical	--	--	--	--	--	--	.12	--	--
Food Service	-.07	--	--	--	--	--	--	--	--
Protective Svc	--	-.08	--	--	-.08	--	--	--	--
Job Preferences									
Security	--	--	--	--	--	.09	--	.12	.09
Autonomy	--	.09	--	-.11	--	--	--	--	--
Routine	-.06	-.11	--	--	--	--	--	.07	--
ADJ, UNCORR R	.560	.305	.464	.352	.591	.401	.481	.507	.294

Table 4-29

Correlations Between the Predictor Constructs and the Army-Wide Criterion
Constructs Combined Across Batch A MOS[a]

Predictor Construct	Criterion Construct				
	General Soldiering (raw score)	Effort and Leadership (resid score)	Effort and Leadership (raw score)	Personal Discipline (raw score)	Phys Fitness Mil Bearing (raw score)
ASVAB Factors					
Technical	.55	.39	.28	.12	-.08
Verbal	.52	.35	.20	.10	-.07
Quantitative	.54	.36	.23	.14	-.01
Speed	.37	.29	.21	.11	.07
Cognitive Constructs					
Overall Spatial	.59	.38	.24	.11	-.03
Computer Constructs					
Complex Perc Speed	-.21	-.17	-.13	-.03	-.04
Complex Perc Accy	.30	.18	.12	.08	-.01
Number Speed/Accy	-.44	-.31	-.21	-.09	-.01
Psychomotor	-.40	-.27	-.20	-.04	-.01
Simp Reaction Accy	.18	.09	.05	.05	-.05
Simp Reaction Speed	-.19	-.13	-.08	-.01	-.06
ABLE Constructs					
Adjustment	.18	.22	.23	.13	.17
Physical Condition	-.03	.09	.10	-.02	.30
Dependability	.09	.15	.21	.30	.22
Achievement	.16	.30	.33	.20	.27
AVOICE Constructs					
Audiovisual Arts	.02	.02	.01	.00	.07
Combat Related	.23	.22	.19	.00	.03
Food Service	-.12	-.14	-.11	-.06	.00
Structural/Machines	.06	.06	.06	-.05	-.01
Protective Services	-.04	.03	.04	-.04	.02
Skilled Technical	.04	.07	.06	.05	.11
Job Constructs					
Autonomy	.13	.15	.09	-.02	-.02
Routine	-.21	-.20	-.15	-.06	-.04
Job Security	.09	.11	.10	.05	.09

[a]Corrected for range restriction.

151

Table 4-30

Correlations Between the Predictor Constructs and Core Technical Proficiency[a]

Predictor Construct	MOS								
	11B	13B	19E	31C	63B	64C	71L	91A	95B
ASVAB Factors									
Technical	.60	.36	.56	.59	.69	.55	.37	.61	.51
Verbal	.63	.33	.49	.67	.50	.44	.56	.71	.59
Quantitative	.60	.32	.49	.67	.45	.46	.63	.64	.59
Speed	.48	.25	.28	.57	.29	.27	.52	.56	.47
Cognitive Construct									
Overall Spatial	.63	.41	.55	.58	.56	.51	.57	.64	.56
Computer Constructs									
Complex Perc Speed	-.33	-.15	-.17	-.25	-.24	-.25	-.11	-.28	-.20
Complex Perc Accy	.38	.24	.32	.22	.16	.28	.40	.25	.26
Number Speed/Accy	-.48	-.30	-.42	-.62	-.37	-.38	-.50	-.57	-.53
Psychomotor	-.43	-.30	-.36	-.34	-.36	-.34	-.26	-.44	-.32
Simp Reaction Accy	.17	.11	.26	.17	.14	.19	.27	.16	.20
Simp Reaction Speed	-.17	-.19	-.15	-.10	-.23	-.19	-.11	-.21	-.23
ABLE Constructs									
Adjustment	.26	.13	.18	.06	.21	.07	.20	.12	.27
Physical Condition	.06	-.04	-.09	-.13	-.13	-.07	-.12	-.09	-.13
Dependability	.16	.01	.09	.04	.00	.01	.21	.18	.24
Achievement	.31	.06	.16	.14	.20	.09	.27	.22	.25
AVOICE Constructs									
Audiovisual Arts	.04	-.05	-.01	.20	-.14	.00	.19	.13	-.14
Combat Related	.23	.21	.31	.08	.31	.24	.02	.22	.03
Food Service	-.30	-.14	-.14	.01	-.20	-.14	-.03	-.09	-.19
Structural/Machines	-.12	.09	.06	.05	.41	.16	-.19	.01	-.19
Protective Svc	-.05	-.08	-.04	-.01	-.10	-.05	.01	-.13	-.16
Skilled Technical	.07	-.03	.09	.12	-.08	.00	.17	.00	-.03
Job Preferences									
Autonomy	.21	.22	.09	.22	.25	.21	.21	.23	.09
Routine	-.27	-.18	-.27	-.19	-.21	-.20	.19	.22	-.30
Job Security	.14	.13	.05	-.02	.06	.14	.20	.18	-.01

aCorrected for range restriction.

The results summarized in this section were impressive and they have formed the basis for modifications to the ASVAB Aptitude Area composites. However, to realize the full benefit of these results, the following things must happen. Both the covariance structures and the estimates of predictive validity must be cross-validated with a genuine predictive design (i.e., the Longitudinal Validation), rules for forming criteria composites must be developed, the utility of accurate predictions must be estimated, the specifics of the full selection/classification/promotion decision system must be modeled, and the effects of using the new predictors in various combinations under a variety of goals and constraints must be evaluated.

A method for obtaining criterion composites and subcomposites has been developed, the utility of a complete set of MOS by performance level combinations has been estimated (presented in Chapter 5), and the data from the Longitudinal Validation sample have been collected. Further work remains on the measurement of second-tour performance and on the full operational model of the complete decision system.

WEIGHTING CRITERION COMPOSITES

The Concurrent Validation results indicated that each of the five criterion components can be predicted with considerable validity and that the validity of the different predictor domains varies systematically across criterion components. A subsequent focus was on the best method for obtaining importance weights when the five components are combined into an overall composite index of performance (Sadacca, Campbell, White, & DiFazio, 1988). Consequently, weighting judgments were gathered from NCOs and officers familiar with each MOS.

The Pilot Experiments

Three pilot experiments were conducted to select the construct weighting procedure. The goal in conducting the experiments was to select one or more construct weighting procedures that would be acceptable to the Army and would yield a reliable, valid set of weights for each of the sampled MOS when the procedures were applied by the appropriate subject matter experts. The experiments and their results will be described briefly prior to describing the actual factor weighting procedure.

The general procedure was that of a small group workshop of 10-16 officers who tried different methods and evaluated the ease of use, acceptability, and perceived validity of each method. The reliabilities and distributional preperties of the assigned weights were also analyzed.

Experiment One

In the first experiment, three procedures were used and all involved direct judgments of the relative weight for each performance construct in forming an overall composite score. In procedure A, the officers were first asked to rank order the constructs and to assign 100 points to the first ranked. The other constructs were scaled so as to produce a ratio estimate.

In procedure B, the officers divided 100 points among the constructs in a manner that reflected the relative weight. In procedure C, 15 pairs of factors were presented in a paired comparison protocol. For the paired comparisons, the order of presentation followed the optimization procedure worked out by Ross (1934), and the officers' task was to divide 100 points between the two constructs being judged in any given pair.

The judgments were made in the context of three different scenarios which described a peacetime condition, a period of heightened tensions, and a wartime setting in which hostilities had just broken out. Each officer used four 7-point scales to evaluate the weighting methods on the following dimensions:

(1) Acceptability to the Army.
(2) Ease of making the judgments called for by the method.
(3) Their confidence in the validity of the judgments made.
(4) The amount of agreement with other workshop participants that could be expected.

After the ratings were completed, an informal discussion period was held to solicit opinions about the methods. The officers generally expressed preference for procedures A and C over procedure B and thought that the time they spent worrying about whether the sum of their weights equaled 100 detracted from their ability to judge the relative importance of the weights. It also seemed that a heightened tension scenario would evoke a more uniform frame of reference across the many different kinds of SMEs providing the MOS construct weights.

Experiment Two

The second pilot experiment used two additional methods, both variants of a conjoint procedure, in two 4-hour workshops. One was attended by 15 officers, the other by 15 NCOs. The three weighting methods are described in the following instructions to the participants:

(1) Rank order the five constructs, assign 100 points to the first ranked construct, and then scale the other constructs accordingly (same as procedure A in Experiment 1).

(2) Based upon their scores on the separate constructs, rank order 25 infantrymen in order of their overall performance. (For each of the infantrymen, a different set of performance scores on the five constructs was given on 7-point scales that range from the lowest level of performance to the highest.

(3) Based upon their scores on two constructs, rank order 10 sets of 13 infantrymen in order of their overall performance. (In each set, the performance scores on two constructs are given on the same 7-point scales used in the second method above. A set of 13 infantrymen is given for each of the 10 possible pairs of the five constructs.

154

The second and third methods are variants of the conjoint approach to scaling. The judges' weights for the performance constructs are inferred from the rank order given sets of hypothetical soldiers whose performance on the constructs has been systematically varied. Both officers and NCOs generally preferred the direct estimation method most and the conjoint full profile method least.

In general, the conjoint paired comparison method yielded the highest intraclass reliability estimates for both the officers and NCOs while the conjoint full profile method had the lowest values. The correlation between the mean officer and NCO weights obtained from the conjoint paired comparisons method also was the highest ($r = .60$). The mean weights obtained from the direct estimation and the conjoint paired comparison methods were highly correlated ($r = .93$) while the correlations of these weights with those obtained from the conjoint full profile method were quite low.

On the basis of these results, it was decided to drop the conjoint full profile method from further consideration.

Experiment Three

The third pilot study also involved two 4-hour workshops, composed of seven officers and eight NCOs. Each participant used the three different weighting methods described below.

Based on scores on two constructs, participants were asked to rank order 21 sets of 13 infantrymen in order of their overall performance. This is the same conjoint paired comparison procedure used in the second experiment, but in addition, the judges assigned overall performance scores that reflected the solders' relative overall performance.

The participants were then asked to rank order the constructs, assign 100 points to the first ranked construct, and then scale the other constructs accordingly (the direct estimation procedure used in Experiments 1 and 2).

The third method was a variant of the second and incorporated a Delphi procedure. Participants first indicated why they had ranked and weighted the performance factors as they had in method 2 above. The reasons were passed around to the other workshop participants. After considering this feedback information, the participants reassigned weights to the performance factors, using method 2 above. The Delphi procedure was then repeated once more.

Several inferences were made from the data. First, there was no evidence that the one-rater reliabilities were improved substantially by adding the requirement to provide overall performance scores in addition to ranks in the conjoint paired comparison method. Nor were agreement indexes improved by adding the requirement to obtain Delphi feedback.

The choice between the direct estimation method and the conjoint paired comparison-ranking method was not clear-cut. The direct estimation method generally received higher evaluation ratings in both Experiments 2 and 3 and

would obviously take less time to administer than the conjoint method. On the other hand, the officer and NCO one-rater reliabilities obtained for the conjoint method were somewhat higher in both experiments. However, both the direct estimation and paired comparison methods had correlations between the officer and NCO mean weights above .80 in both experiments. The correlations between the mean weights obtained in Experiment 2 with those obtained in Experiment 3 were very high for both methods (.96 for the direct estimation and .97 for the conjoint method).

In short, both appeared to be sound methods and it was decided to use both to obtain the construct performance weights for the Project A MOS sample.

Obtaining the Performance Construct Weights

The component weights were collected in a series of 2-hour workshops. Separate workshops were held for NCOs and officers at each of two posts for each MOS. Of a total of 36 judges for each MOS[3], half were to come from field units (FORSCOM and USAREUR) and half from proponent posts (TRADOC). The judges were to be evenly divided among NCOs, company grade officers, and field grade officers. Table 4-31 shows the total sample of 702 judges subdivided by MOS, type of post, and grade level. Although some individual MOS proportions did not meet the target, overall the proportions of officers to NCOs and judges from field units to proponent MOS posts were close to the desired composition.

At each workshop, after a briefing on Project A, the participants were first given general instructions which covered the background and purpose of the workshop, and descriptions of the performance components (constructs) and the two methods (direct estimation and conjoint paired comparison-ranking) that would be used to obtain weights for the components. The components to be weighted were the five job performance criterion factors developed as part of Project A's performance modeling effort. The two scaling methods were then administered, always in the same order.

Analysis and Results

To better reflect the combined judgments of the construct weights across the judges for each MOS, the data from each judge were standardized prior to averaging. For the direct estimation method, the average of the five construct weights of all judges was set at 20.0, and the average of the five weights for any group of judges within and across MOS was also set at 20.0. The mean weight of a given construct obtained by averaging the judges' individual weights could, of course, be different from 20.

---

[3]MOS 96B (Intelligence Analyst), which was added to the LV MOS sample to improve job area coverage, was included in these workshops, making a total of 20 MOS studied in this effort.

Table 4-31

Composition of Judging Sample* for Weighting Project A MOS

	MOS	Type of Unit				Total	
		Field		Proponent			
		Officer	NCO	Officer	NCO	Officer	NCO
11B	Infantryman	17	6	19	6	36	12
12B	Combat Engineer	17	4	12	6	29	10
13B	Cannon Crewman	6	6	21	6	27	12
16S	MANPADS Crewman	11	6	11	5	22	11
19E	Armor Crewman	11	5	14	6	25	11
27E	TOW/Dragon Repairer	--	6	16	5	16	11
31C	Single Channel Radio Oper	13	6	12	6	25	12
51B	Carpentry/Masonry Specialist	4	6	27	6	31	12
54E	Chemical Operations Spec	20	14	--	-	20	14
55B	Ammunition Specialist	4	3	24	9	28	12
63B	Light Wheel Vehicle Mechanic	7	2	20	11	27	13
64C	Motor Transport Operator	10	5	12	6	22	11
67N	Utility Helicopter Repairer	12	1	17	12	29	13
71L	Administrative Specialist	13	6	9	7	22	13
76W	Petroleum Supply Specialist	10	11	--	--	10	11
76Y	Unit Supply Specialist	15	5	8	5	23	10
91A	Medical Specialist	25	13	--	--	25	13
94B	Food Service Specialist	12	7	8	4	20	11
95B	Military Police	23	13	--	--	23	12
96B	Intelligence Analyst	--	--	11	6	11	6
		230	125	241	106	471	231

*In addition to the 702 officers and NCOs listed in this Table, there were
10 judges whose grades were unknown, making the total sample 712.

For the conjoint method, the data from each judge was scaled using a
method developed by Comrey (1950) which is described in Torgerson (1958).
Essentially, the multiple regression equation predicting the judge's rank
orders of the two performance construct scores of the 15 hypothetical soldiers
was first obtained for each of the 10 sets of soldiers. The ratio of the two
regression weights for each pair of constructs then became the basic data
entering into the scaling procedure. Since the correlation between the two
construct scores of the 15 hypothetical soldiers on each performance rating
sheet was specified to be zero, the ratio of the regression weights is
directly proportional to the correlation of each set of construct scores with

157

the judge's rank order of the soldiers. The means and standard deviations of the construct scores were equal for all constructs.

The scaling procedure employs a least squares solution to obtain a set of weights that best fit the observed ratios. The resultant weights are so scaled that their geometric mean is 1.0. To facilitate the comparison of the conjoint weights to those obtained by the direct estimation method, the conjoint weights for each judge were also linearly transformed so that their sum was equal to 100 and their average equal to 20.0.

Interjudge Reliability and Intermethod Agreement

The NCO 1-rater and n-rater reliabilities for the direct estimation and conjoint scaling methods were .132/.425 and .153/.509, respectively. The corresponding values for officers were .278/.864 and .287/.867.

The correlations across the 20 MOS of the average weights derived from the direct estimation and conjoint scaling methods using officer judgments ranged from .836 to .996; the average intermethod agreement was .951. The corresponding range for the NCOs was .017 to .922 and their average MOS intermethod agreement was .653. These intermethod results reflect in part the lower 1-rater reliabilities obtained for the NCOs under both methods; also, there were fewer NCO judges.

Comparison of the Direct Estimation and Conjoint Scaling Methods

To decide whether the final sets of weights should be obtained from the direct estimation or the conjoint method, the two sets of weights were compared on several indexes. Though the differences were in general slight, they all favored the conjoint method. The 1-rater and n-rater intraclass reliabilities for the combined group of officers and NCOs tended to be slightly higher for the conjoint method across the 20 MOS. While the differences between the reliabilities for the two scaling methods were slightly greater for the NCOs than for the officers, the difference favored the conjoint method in each case.

Also, the weights assigned the constructs by the NCOs correlated higher with those assigned by the officers when the conjoint scaling method was used.

The Final Weight Estimates

Considering the above findings, the decision was made to favor the weights derived from the conjoint scaling method in combining the individual construct scores into an overall composite measure of performance. They are shown in Table 4-32.

It should be borne in mind that the weights are based on comparative judgments of the constructs within each MOS and should not be used for comparisons of importance across MOS. Nevertheless, it is interesting to note whether the relative pattern of weights differ across MOS and whether some constructs are fairly consistently given relatively higher weights than others.

158

Table 4-32

Mean Construct Weights by Grade and MOS: Conjoint Method

MOS	MOS Skills			General Skills			Ex. Leadership			Main. Discipline			Military Bearing		
	NCO	Off	Total	NCO	Off	Total	NCO	Off	Total	NCO	Off	Total	NCO	Off	Total
11B	19.5	22.9	22.0	18.7	18.5	18.5	21.7	29.1	27.3	22.2	17.2	18.4	18.0	12.3	13.7
12B	20.2	18.4	18.8	22.8	19.7	20.4	21.1	30.2	28.1	23.8	20.3	21.1	12.1	11.5	11.6
13B	28.6	22.7	24.1	16.2	19.2	18.5	21.1	27.7	26.2	20.7	18.3	18.9	13.3	12.1	12.4
16S	20.8	25.9	24.6	16.8	16.3	16.4	23.1	26.3	25.5	24.8	20.2	21.3	14.5	11.4	12.1
19E	30.0	29.4	29.6	20.2	21.1	20.8	16.5	20.5	19.3	23.3	17.9	19.5	10.0	11.0	10.7
27E	21.7	24.2	23.5	20.9	18.0	18.8	21.9	22.4	22.3	21.0	23.0	22.4	14.5	12.4	13.0
31C	29.4	29.0	29.1	20.0	20.3	20.3	16.7	22.0	20.8	19.0	17.3	17.6	14.9	11.4	12.2
51B	19.4	25.6	23.9	20.2	17.2	18.0	22.7	25.6	24.8	20.2	19.7	19.8	17.5	11.9	13.4
54E	28.6	25.4	26.5	16.7	21.5	19.8	17.3	20.7	19.5	20.6	19.8	20.1	16.8	12.6	14.1
55B	32.4	22.4	24.9	18.9	19.5	19.4	14.8	27.8	24.7	16.6	19.5	18.8	17.3	10.8	12.3
63B	21.4	27.5	25.6	21.4	18.1	19.1	21.8	23.5	23.0	19.4	21.1	20.5	16.1	9.9	11.8
64C	21.8	26.1	24.8	16.9	22.8	20.9	23.2	21.8	22.2	20.9	15.4	17.1	17.2	14.0	15.0
67N	28.7	25.9	26.7	15.8	15.9	15.9	23.7	25.3	24.9	20.7	22.2	21.8	11.1	10.6	10.6
71L	20.9	24.1	23.1	20.4	19.9	20.1	21.2	22.7	22.3	20.7	21.0	20.9	16.8	12.3	13.7
76W	29.4	23.6	25.7	19.1	17.2	17.9	17.3	25.0	22.2	18.5	22.9	21.3	15.7	11.4	12.9
76Y	26.3	25.7	25.8	23.3	21.7	22.1	18.7	19.8	19.5	15.8	17.5	17.1	15.8	15.3	15.4
91A	23.2	26.9	27.3	18.8	16.6	17.3	20.0	23.1	22.1	21.3	22.5	22.1	11.7	11.0	11.2
94B	19.9	24.5	23.0	15.0	17.4	16.6	26.7	26.2	26.4	23.1	20.9	21.6	15.3	11.0	12.4
95B	17.5	20.0	19.2	23.5	27.8	26.5	26.4	20.5	22.4	22.2	19.1	20.1	10.3	12.6	11.8
55B	18.9	28.7	25.2	22.2	18.8	19.6	19.0	23.3	21.8	25.2	19.8	21.7	14.7	10.0	11.7
AVG	24.2	24.9	24.7	19.4	19.3	19.3	20.8	24.2	23.3	21.0	19.8	20.1	14.7	11.8	12.6

159

For all 20 MOS, Physical Fitness/Military Bearing received the lowest relative weight. In 13 of the 20 MOS, Core Technical Skills received the highest relative weight, while the Effort/Leadership construct was second overall. The Effort/Leadership component received the highest relative weight in 6 of the 20 MOS. For the most part, the Core Technical construct received the highest weight for the technical MOS in the sample and the Effort/Leadership construct received the highest weight for the combat MOS. The General Soldiering construct received the highest weight for only one MOS, Military Police (95B). These MOS differences in the constructs receiving the highest weights undoubtedly contributed to the significant Construct by MOS interaction.

Significant mean differences between the weights assigned by the officers and NCOs were found for two constructs: Officers gave significantly higher relative weights to the Effort/Leadership construct than did NCOs, while NCOs gave higher weights to the Physical Fitness/Military Bearing construct than did officers. The NCOs may have been giving relatively more weight to aspects of first-tour soldiers' performance that were of more immediate concern to them. Although the mean differences were only significantly different at the .10 level, the NCOs gave the Personal Discipline construct weights that were higher on the average than those assigned by the officers.

Summary

The five Project A performance constructs received significantly different patterns of weights in different MOS and the different groups of experts agreed, in general, on the relative ranking of the weights. For example, the Effort/Leadership construct tends to be rated highest among the combat MOS.

Multiple judges per MOS, about 30 on the average, produced n-rater reliabilities that are quite respectable (above .95 for most MOS). The high intermethod correlations (about .95 on the average) between the construct weights obtained by the direct estimation and conjoint methods for the separate MOS further document the reliability of the means of the scaled weights.

That different groups of judges may provide somewhat different MOS weights can be seen in the relatively low correlations between the officer and NCO weights. The NCOs tended to give relatively higher weights to the Physical Fitness/Military Bearing construct, while the officers attached more importance to the Effort/Leadership construct.

Though there were statistically significant differences in the mean weights assigned under the three scenarios, the very small differences will have little impact on the relative ranking of soldiers on the overall performance composites for an MOS. A more critical question is how much impact will the weights themselves have? That is, would a different set of predictors be selected using a weighted composite for validation than would have been selected if the constructs had been weighted equally? And perhaps, even more

160

importantly, would different classification assignments be made as a result of using the scaled weights?

The answers to these questions obviously depend not only on the set of weights used but on such factors as the intercorrelations among the construct performance scores, the validity of the predictor battery, the amount of differential prediction it affords across Army jobs, the MOS selection standards in effect, and the assignment algorithms employed. The most feasible way to address these issues is through a series of sensitivity analyses that portray the effects of these parameters on selection and classification validity. These analyses remain to be done.

Chapter 5
SCALING THE UTILITY OF INDIVIDUAL PERFORMANCE

Finding a way to place value on different levels of job performance across different MOS was one of the research objectives for Project A. Two principal factors made it difficult to apply previous civilian research on utility metrics and utility estimation to the Army context. First, compensation practices in the Army are quite different than in the civilian sector. Salaries do not differ by MOS and thus cannot be used as an index of a job's relative worth to the organization. Second, the Army is not in business to provide products or services so as to maximize profit. Its overall mission to be prepared to defend the United States against external military threats makes it inappropriate to put a monetary value on success or failure or to think of the utility of jobs in terms of their monetary benefit. Thus dollars may not be an appropriate metric with which to evaluate a new classification system aimed at maximizing preparedness for catastrophic events. Nevertheless, military resources are not unlimited. Choices among alternative personnel practices will have to be made, whether or not there is an explicit utility metric on which to make comparisons.

The utility problem for Project A was one of assigning utility values to MOS-by-Performance-Level combinations. That is, if it is true that personnel assignments will differ in value to the Army depending on the specific MOS to which an assignment is made and on the level at which an individual will perform in that MOS, then a classification strategy that has a validity significantly greater than zero will increase in value to the extent that the differential values (utilities) can be estimated and made a part of the assignment system.

The problem of estimating such utility values was composed of a number of specific questions: How should performance levels be defined? Should it be in terms of general performance defined only as relative level (e.g., percentiles), with behavioral anchors developed by means of critical incident methodology? Or should individual performance components be defined and then explicitly weighted for combination into a total score? What is the most appropriate metric for describing the relative value of differential assignments? Since the dollar metric seemed not to be appropriate for the Army context, this was a very difficult issue for Project A. It required an exploratory approach.

What method(s) should be used to estimate utility? Only two options seemed even possible. First, it might be possible to relate the performance of individuals to some kind of "bottom line" measure that Army management would consider an appropriate metric, such as realistic field exercises. The difficulties with this approach revolve around feasibility, expense, and the necessity for equating scores in some way across MOS. A second alternative was to appeal to scaling technology and use expert judges to estimate the relative value of differential personnel assignments, and this is the course that was followed.

The general procedure used in Project A to obtain utility values for different levels of predicted performance in each MOS was divided into three phases. Phase one was exploratory in nature and intended to uncover the major issues. The goal of phase two was to evaluate alternative scaling methods and

develop the procedure to be used. In phase three the selected methods were used to obtain the final scale values. (See Sadacca, White, Campbell, DiFazio, & Schultz, 1988.)

PHASE ONE: EXPLORING ISSUES

Phase one consisted of a series of seven small group workshops with Army officers. Each workshop was divided into a period for trying out prototypic judgment tasks and a period for open-ended discussion of issues. These questions were used to guide the discussions:

(1)   How shall measures of performance be weighted and overall performance defined?

(2)   What kinds of scaling judgments can officers reasonably be asked to make?

(3)   Are there major scenario effects on performance factor weights and utility judgments?

(4)   In what metric should the utility of enlisted personnel assignments be expressed?

(5)   What is the form of the relationship between performance and utility within MOS?

(6)   Who will make the best judges for the final scaling?

The prototypic judgment tasks that were tried out in phase one were of the following general nature:

(1)   Assignment of importance weights to performance factors.

(2)   Rank ordering of overall utility of MOS x Performance Level combinations when performance was defined in percentile terms.

(3)   Ratio judgments of comparative utility for different MOS x Performance Level combinations.

The specific reactions of each participant to the sample scaling tasks were also used as items for general discussion.

Perhaps the most significant finding was that Army officers would be willing and able to assign differential utility values across MOS and performance levels. When asked their reaction to expressing the differential worth or utility of soldiers in terms of dollars, the officers in the workshops reacted very negatively to this concept, citing possible adverse political consequences as well as internal Army morale problems if dollar figures were placed on soldiers' worth.

Perhaps the next most significant finding was that fairly stable scale values could be obtained from averaging across a relatively small number of officer/judges. In these exploratory trials there was considerable agreement across workshops on the scale values assigned to selected MOS x Performance Level combinations. Judges seemed to have a common frame of reference

concerning what different performance levels meant; and, in the absence of any specification, everyone imposed the same scenario or context (i.e., being prepared for a major conflict in Europe).

The workshop groups also agreed that the scenario(s) used should be free of the detail that suggests greater or less utility for certain specific MOS. An acceptable metric for expressing utilities of soldiers in wartime would be the utility of a 50th percentile Infantryman (his value for the survival of the unit and in replacing troop losses is much more readily apparent). Directions to the judges should be reassuring concerning inconsistencies that may occur in a long series of judgments.

PHASE TWO: EVALUATING METHODS

The second phase was devoted to developing and evaluating the final procedures to be used in assigning utilities to performance levels in all entry-level MOS. Several inferences were made from the exploratory findings in earlier workshops. First, the apparent nonlinear relationships between utility and performance found in some MOS would necessitate obtaining judgments of the utility of at least five performance levels within each MOS. Five data points would allow the derivation of a best fitting utility/performance curve with two inflection points (if necessary) within an MOS. Second, assigning utility scale values to at least five performance levels in 276 MOS was much too onerous to assign to any one judge. Third, high correlations between different methods suggested that a combination of methods might allow the total scaling task to be accomplished more efficiently.

The goal was to place all 276 x 5 MOS/performance level combinations on the same ratio scale, which would permit utilities to be summed across individual MOS assignments in comparing selection/ classification systems. Consequently, an additional 12 workshops were conducted with small groups of officers to try out various scaling methods. These included rank ordering, paired comparisons, a conjoint scaling procedure, the sorting or placement of MOS/performance level combinations into piles (i.e., a Thurstone sort), and the direct estimation of ratio scale values using a standard MOS/performance level set a. 100. Of these techniques, the last two were the scaling procedures eventually selected.

The rank ordering procedure produced much negative reaction because of the time it took, the inability to assign ties, and the requirement to rank some MOS at the very bottom.

A major change during phase two involved placing the judgments in a selection and classification context. That is, the instructions were changed to ask for judgments of the utility of predicted performance of Army applicants or recruits rather than actual performance of incumbents (as had been the case in earlier workshops). The judges were asked to assume that the performance percentiles given were accurate estimates of future on-the-job performance percentiles if the applicants or recruits were actually assigned to the MOS. After this adjustment was made, none of the judges in subsequent workshops objected to the basic concept of assigning differential utilities to various MOS/performance levels.

165

Two variants of the method of paired comparisons were also tried out using a limited number of MOS/performance level combinations. However, the methodology was time consuming and the officers felt they should be allowed to indicate that some applicants should not be selected at all. The judgment was subsequently shifted from predicted performance levels of applicants to that of recruits (selected applicants), thereby eliminating the "do not select" alternative.

To divorce both troop strength and troop replacements from utility judgments, judges were told that the field strength of all MOS was 70 percent and that the problem of compensating for troop losses was being handled by another part of the assignment algorithm and should not enter into their judgments.

A conjoint scaling method was also tried out but the method was much too difficult and time consuming for use in scaling this number of stimuli.

One method that did prove effective for making large numbers of scaling judgments was the pile placement method in which the judges sorted cards containing MOS/performance level combinations into piles, based upon their perceived utility or selection priority. Seven piles of predicted performance utility were used, ranging from negative through zero utility to high utility. The judges initially sorted 135 MOS/performance level combinations, then 210 combinations, and eventually 280 combinations without complaining about the judgment burden.

Likewise, the ratio judgment method, in which judges evaluated MOS/ performance level utilities in relationship to that of a 90th percentile Infantryman, was stepped up to 60 combinations without becoming burdensome. The one-rater intraclass correlation reliability estimate for the pile placement procedure was .58 and the comparable coefficient for the direct ratio judgment was .65. These results indicated that satisfactory reliabilities for mean utilities could be obtained by both methods if the means were based upon 10 or more judges. The correlation between the mean utilities assigned by the 12 officers to the 60 common combinations, using the two methods, was .89.

Considering all the information available from the first and second phase workshops, Project A staff decided to use the pile placement and direct ratio estimation methods in the final determination of the utilities of approximately 276 MOS x 5 performance levels, or 1,380 combinations. The pile placement method provided a means of reliably scaling the utility of large numbers of combinations on an interval scale in a reasonable time period, while the direct estimation method could be used to place a limited number of combinations on a ratio scale having a meaningful zero point. If a set of stimuli (MOS x Performance Level combinations) was scaled by both methods, the data could be used to develop an algorithm for estimating ratio scale values from interval scale values.

PHASE THREE: OBTAINING A COMPLETE SET OF UTILITY ESTIMATES

The results of the exploratory workshops were largely successful. Utility scale values varied across MOS in a manner generally consistent with expectations, and interjudge agreement was high enough to indicate that fairly stable scale values could be obtained by averaging across officer judgments.

166

The next goal was to assign a utility to any predicted level of performance for any entry-level MOS such that the values could be used to (a) make classification decisions and (b) assess the net gain to the Army of using new selection/classification procedures.

Procedures

The scaling task considered all MOS that required an ASVAB Aptitude Area score for assignment; that is, 276 MOS times 5 levels or 1,380 MOS/performance level combinations to be judged separately. To make the scaling task more acceptable to the judges, seven separate sets were used. The first set of 12 MOS times 5 performance levels, or 60 combinations, was to be judged by all judges as the basis for a common scale. The remaining 264 MOS were grouped into six comparable subsets of 44 MOS each. Each deck thus contained 280 MOS/performance level combinations--12 common plus 44 noncommon MOS times 5 performance levels.

Sample of Officers Used as Judges

To ensure a total sample of 60 officers (10 officers x 6 decks) utility workshops were held at 6 CONUS Army posts and in USAREUR. Altogether, 74 field grade officers attended the workshops--54 majors, 13 lieutenant colonels, and 7 colonels.

The Utility Judgment Workshops

After a brief overview of Project A, a description of the agenda, and completion of a Background Information Sheet, the leader discussed three critical assumptions:

(1) The military context is a period of heightened tensions with an increasing probability that hostilities will break out in Europe, Asia, the Caribbean, Latin America, and Africa. Some potential enemies have nuclear and chemical capability and air parity does exist.

(2) The overall MOS performance measure for each MOS represents an optimally weighted combination of multiple performance factors.

(3) The predicted performance levels for the recruits are accurate. That is, the recruits will actually perform at the predicted levels.

For the pile placement method, the judges were to sort the MOS/performance level combinations into one of seven piles ranging from positive, to zero, to negative utility. For the direct judgment method, the participants wrote the value, 100, on the 90th percentile Infantryman card, and then assigned a utility value to each of the remaining 59 MOS/ performance level combinations so as to establish a utility ratio using the 90th percentile 11B as the standard. Zero and negative utility values were permitted.

167

Analyses

Reliability and Validity Analyses

After extensive outlier analysis, seven extremely atypical judges were removed and all reliability and validity analyses, as well as the utility value estimates, were based on the remaining 67. The n-rater reliabilities for the six separate decks (based on an n of about 11 judges on the average) ranged from .958 to .976 for the pile placement data. The n-rater (67 judges) reliability for the direct judgment utilities of the common combinations was .992. The corresponding reliability for the pile placements of the common combinations (across all decks and the 67 judges) was .995. The correlation obtained between the average scale values from the two methods across the 60 common combinations was .98.

This high correlation was not wholly attributable to judges simply agreeing that good performance is worth more than poorer performance. This can be seen by the correlations between average pile placement and direct judgment utilities attained when the correlations are computed across the 12 common MOS holding the performance percentile constant. These correlations had an average value of .77. The n-rater (67 judges) reliabilities averaged .89 and .82 respectively for the pile placement and direct judgment utilities, when the reliabilities were computed for each percentile level separately.

Comparison of Utility Ratings by Different Officer Specialties

Analyses were conducted to determine whether officers in different military primary specialties assigned significantly different utilities to the common MOS/performance level combinations. In all, only 10 of the more than 250 statistical tests run were significant at the .05 level. Examination of the significant differences that were obtained did not reveal any trend in the data indicating that certain types of officers favored particular MOS or performance levels.

Estimates of Ratio Scale Utilities From Pile Placement (Interval) Data

A basic objective of the overall research design was to place all 1,380 MOS/performance level combinations on the same utility scale. Using the averages (across all judges) of the direct judgment utilities assigned the 60 common combinations as the dependent variable, and the pile placement of the same common combinations as the basic independent variable, an equation was derived expressing direct judgment utilities as a function of average pile placement. This equation was then used to estimate the ratio scale values (direct judgment utilities) for each group of judges.

Alternative regression equations as estimates of ratio scale utilities from the pile placement data were evaluated on a hold-out sample of 20 combinations. The overall multiple correlations were very high, .97 on the average, although in general the equations tended to underestimate the utilities of the hold-out combinations having high actual utilities, and slightly overestimate the utilities of the combinations having low actual utilities. The best balance was achieved by using average pile placement and both its square and cube as the independent variables. The sign of the weights obtained formed a fairly consistent pattern with average pile placement always

168

having a positive weight, and the square and cube of average pile placement having negative and positive weights, respectively.

Cross-Validation of Estimation Equations on a Hold-Out Sample

The ten participants in the last utility workshop were given an additional 40 combinations (8 MOS x 5 levels) on which to make their direct judgments of utility. The means cf the direct judgment utilities given these 40 combinations were estimated by formulas derived for each deck, excluding any of the data obtained from the last workshop.

Very high correlations (.97) were again obtained between the utilities estimated from the separate deck equations and the hold-out sample direct judgment utilities. Moreover, the direct judgment means, standard deviations, and ranges for the 40 extra combinations obtained from the hold-out were quite similar to those estimated from the equations.

THE FINAL UTILITY VALUES

The analyses supported the conclusions that:

(1) For both methods the reliability of a single judge is reasonably high.

(2) For both methods the reliability of the average value produced by 11 judges or more is *very* high.

(3) Reliabilities are high even when performance level is controlled and differences are due only to MOS differences within performance level.

(4) The agreement between the two utility scaling methods is very high and equal to the limit of their reliabilities.

(5) Judges from different posts or MOS backgrounds do not produce different patterns of scale values.

(6) A relatively simple exercise in equation fitting produced a useful method for estimating ratio scale values (which could not be obtained for all MOS x Performance Level combinations) from the interval scale values which were obtained from all MOS x Performance Level combinations using the pile placement (Thurstone sort) method.

(7) As determined on a cross-validation sample of stimuli, the equations used to estimate ratio values from interval data were highly accurate ($R_{estimated}$ x actual = .97).

The derived equations for each deck were used to estimate the ratio scale utilities for the noncommon MOS/performance level combinations. These values represent the bottom line of the Project A utility scaling work. Within the limits of the reliability and validity evidence discussed here, the 1,365 combinations have been placed on the same ratio scale.

169

DISCUSSION AND CONCLUSIONS

The assigned utilities had very high reliabilities and the estimated ratio scale values correlated very highly with direct judgments. These results held even when performance level was held constant. A personnel assignment algorithm that took into account the value of performance would most likely be able to effect more optimal Army-wide assignments than one that did not.

However, a number of problems need to be addressed before utilities similar to the ones obtained in this research can be used operationally. One problem concerns the optimal distribution within MOS, considering both within- and between-MOS utilities as well as the available recruit pool and the quality of existing personnel. This is the issue of average vs. marginal utility (Nord & White, 1988, 1990). Another issue concerns the duration of time that the recruits actually remain in the Army and how to aggregate values over time.

Clearly, this research has affirmatively answered the question of whether a coherent, reliable set of relative utility values could be derived for all performance levels in all entry-level Army MOS. The next steps involve how to make best use of that finding in improving the Army's selection, classification, and assignment processes.

COMPLETION OF LONGITUDINAL VALIDATION PREDICTOR AND END-OF-TRAINING DATA COLLECTION

Since one goal of the LV data collection was to administer the predictors as closely as possible to the point where they would ultimately be administered operationally, testing during Reception Station processing was chosen as the most feasible method of obtaining the desired sample. Soldiers in the LV sample would then be followed into their first tour, where the first-tour job performance measures would be administered, and eventually into their second tour, where the second-tour performance measures could be administered. This data collection process is summarized schematically in Figure 6-1.

LONGITUDINAL VALIDATION

Data Collection Schedule

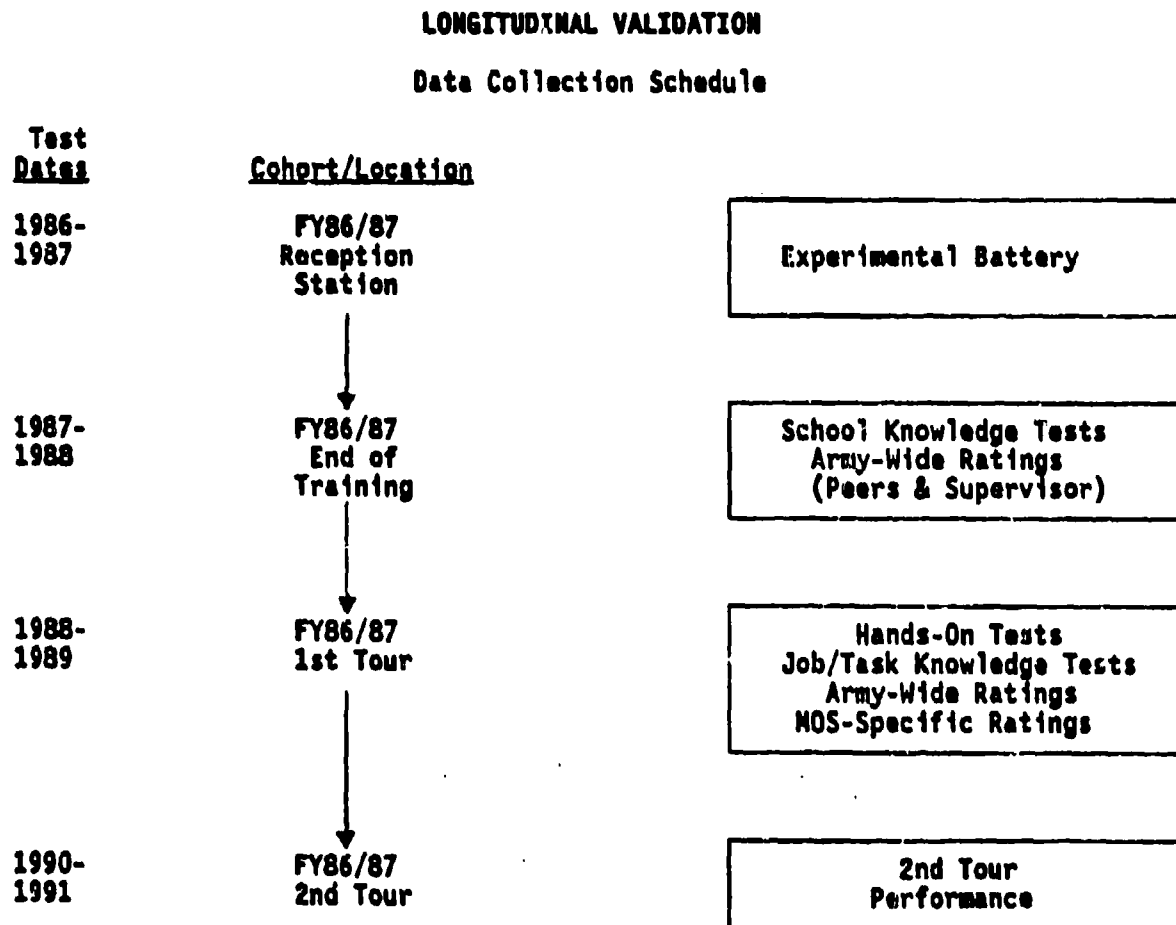| Test Dates	Cohort/Location	
1986-1987	FY86/87 Reception Station	Experimental Battery
1987-1988	FY86/87 End of Training	School Knowledge Tests Army-Wide Ratings (Peers & Supervisor)
1988-1989	FY86/87 1st Tour	Hands-On Tests Job/Task Knowledge Tests Army-Wide Ratings MOS-Specific Ratings
1990-1991	FY86/87 2nd Tour	2nd Tour Performance

Figure 6-1. Longitudinal Validation data collection scheme.

SAMPLE AND SCHEDULE

The sample of MOS for the Longitudinal Validation is shown in Table 6-1. To improve coverage of MOS job families, two MOS (29E, Electronics Repairer, and 96B, Intelligence Analyst) had been added to the sample used for Concurrent Validation and one MOS (76W, Petroleum Supply Specialist) had been deleted. In addition, MOS 19K (M1 Armor Crewman) was added because MOS 19E (M60 Armor Crewman) was being severely scaled back. These modifications resulted in an LV sample of 21 MOS, compared to 19 MOS during the CV phase.

Table 6-1

Project A MOS in Longitudinal Validation Sample

Batch A		Batch Z	
MOS		**MOS**	
11B	Infantryman	12B	Combat Engineer
13B	Cannon Crewman	16S	MANPADS Crewman
19E	M60 Armor Crewman	27E	Tow/Dragon Repairer
19K	M1 Armor Crewman	29E	Electronics Repairer
31C	Single Channel Radio Operator	51B	Carpentry/Masonry Specialist
63B	Light-Wheel Vehicle Mechanic		
71L	Administrative Specialist	54E	NBC Specialist[b]
88M	Motor Transport Operator[a]	55B	Ammunition Specialist
91A	Medical Specialist	67N	Utility Helicopter Repairer
95B	Military Police	76Y	Unit Supply Specialist
		94B	Food Service Specialist
		96B	Intelligence Analyst

[a] MOS 88M was previously identified as MOS 64C.
[b] MOS 54E subsequently became MOS 54B.

To obtain a large enough sample for the extended testing involved in the Longitudinal Validation, each of the eight Reception Battalions was asked to test all Regular Army soldiers entering any one of the 21 MOS listed in Table 6-1 for an entire year. Testing sites and data collection periods were as follows:

Site	Predictor Testing Period
Fort Sill	20 Aug 86 - 20 Aug 87
Fort Benning	27 Aug 86 - 27 Aug 87
Fort Bliss	4 Sep 86 - 4 Sep 87
Fort Knox	10 Sep 86 - 10 Sep 87
Fort McClellan	17 Sep 86 - 17 Sep 87
Fort Dix	24 Sep 86 - 24 Sep 87
Fort Leonard Wood	1 Oct 86 - 1 Oct 87
Fort Jackson	19 Nov 86 - 19 Nov 87

172

THE EXPERIMENTAL BATTERY

Table 6-2 shows the complete array of tests and inventories in the Experimental Battery, the number of items in each, and the time limit (for the timed tests) or approximate time to finish (for the computer-administered tests and the untimed inventories).

Table 6-2

Description of Tests in Experimental Battery

Cognitive Paper-and-Pencil Tests	Number of Items	Time Limit (minutes)
Reasoning Test	30	12
Object Rotation Test	90	7.5
Orientation Test4		10
Maze Test	24	5.5
Map Test	20	12
Assembling Objects Test	36	18

Computer-Administered Tests	Number of Items	Approximate Time
Demographics	2	4
Reaction Time 1	15	2
Reaction Time 2	30	3
Memory Test	36	7
Target Tracking Test 1	18	8
Perceptual Speed and Accuracy Test	36	6
Target Tracking Test 2	18	7
Number Memory Test	28	10
Cannon Shoot Test	36	7
Target Identification Test	36	4
Target Shoot Test	30	5

Non-Cognitive Paper-and-Pencil Inventories	Number of Items	Approximate Time
Assessment of Background and Life Experiences (ABLE)	199	35
Army Vocational Interest Career Examination (AVOICE)	182	20
Job Orientation Blank (JOB)	31	5

The information obtained from Concurrent Validation data analysis was used to make the final revisions to the predictor battery for the Longitudinal Validation. Since the battery had already been through several iterations of data collection, analysis, and revision, the revisions were not substantial.

Of the six cognitive tests, only one had actual item content change. The Assembling Objects test was made more difficult by adding four new items and revising three existing items; two minutes were added to the time limit. For the computerized portion of the battery, minor modifications were made to the instructions, several changes were made in the software, and several items on the Target Identification Test were revised to balance the item types better.

The ABLE revisions included deleting 10 items, revising 16 items, and using a separate answer sheet for responding. For the AVOICE, several changes were made in the scoring procedures, switching already existing items to scales where their item-total score correlations were higher, and in two cases combining two pre-existing scales. Ten items were dropped from the AVOICE, 16 were added, several scales were renamed, and a separate answer sheet was prepared. The JOB was shortened by seven items and had five items reworded, and all scales were reconstituted and renamed, based on factor analyses of the CV data. A list of the scales on all three non-cognitive inventories appears as Table 6-3.

TRAINING PERFORMANCE MEASURES

As part of the Longitudinal Validation, criterion measures of training performance were collected on each individual at the end of AIT or OSUT. The end-of-training measures were administered to soldiers at the eight predictor testing installations and at six other AIT-only installations where the Project A MOS were trained. These 14 installations, the MOS tested at each, and the data collection period for each are shown in Table 6-4.

The training measures consisted of a number of rating scale evaluations collected from the individual's Drill Instructor and the training achievement test previously developed for each MOS.

The development and field testing of the paper-and-pencil achievement tests were described in the FY85 Annual Report (Campbell, 1987a) and in Davis, et al. (1986). The rating scale measures were modified versions of the Army-wide BARS scales used as job performance measures (Pulakos & Borman, 1986). The following scales were used:

A. Technical Knowledge/Skill
B. Effort
C. Following Regulations and Orders
D. Military Appearance
E. Physical Fitness
F. Self-Control
G. Leadership Potential

174

Table 6-3

ABLE, AVOICE, and JOB Scales in Experimental Battery

ABLE Scales

Adjustment: Emotional Stability

Dependability: Nondelinquency
                Traditional Values
                Conscientiousness

Achievement: Work Orientation
              Self-Esteem

Surgency (Leadership/Potency): Dominance
                                Energy Level

Agreeableness/Likability: Cooperativeness

Locus of Control: Internal Control

Physical Condition: Physical Condition

Response Validity Scales: Unlikely Virtues (Social Desirability)
                           Self-Knowledge
                           Non-Random Response
                           Poor Impression

AVOICE Scales

Realistic: Mechanics   Fire Protection
            Heavy Construction   Audiographics
            Electronic Communication Rugged Individualism
            Drafting   Firearms Enthusiast
            Law Enforcement   Combat Vehicle Operator

Conventional: Clerical/Administrative Food Service--Professional
               Warehousing/Shipping   Food Service--Professional

Social & Enterprising: Leadership/Guidance

Investigative: Medical Services   Science/Chemical
                Mathematics   Computers

Artistic: Aesthetics

JOB Scales

Job Pride   Job Autonomy
Job Security   Job Routine
Serving Others   Ambition

Table 6-4

End-of Training Data Collection Sites and Data Collection Period

Site	MOS	End-of-Training Testing Period	
Fort Sill	13B	15 Nov 86	21 Nov 87
Fort Benning	11B	12 Nov 86	4 Dec 87
Fort Bliss	16S	8 Jan 87	22 Jan 88
Fort Knox	19E	6 Dec 86	12 Dec 87
	19K	16 Dec 86	12 Dec 87
Fort McClellan	54B	28 Mar 87	16 Apr 88
	95B	24 Jan 87	16 Jan 88
Fort Dix	63B	7 Mar 87	27 Feb 88
	88M	24 Jan 87	23 Jan 88
	94B	7 Feb 87	4 Feb 88
Fort Leonard Wood	12B	17 Jan 87	9 Jan 88
	51B	31 Jan 87	23 Jan 88
	63B	7 Mar 87	6 Feb 88
	88M	14 Feb 87	30 Mar 88
Fort Jackson	63B	2 May 87	16 Apr 88
	71L	15 Apr 87	6 Apr 88
	76Y	28 Mar 87	2 Apr 88
	94B	18 Apr 87	2 Apr 88
Redstone Arsenal	27E	10 Mar 87	21 Apr 88
	55B	11 Dec 86	3 Mar 88
Fort Lee	76Y	10 Jan 87	17 Feb 88
	94B	10 Jan 87	18 Feb 88
Fort Rucker	67N	17 Jan 87	13 Feb 88
Fort Sam Houston	91A	19 Feb 87	30 Mar 88
Fort Gordon	29E	27 Apr 87	14 Apr 88
	31C	13 Feb 87	18 Apr 88
Fort Huachuca	96B	14 Apr 87	11 Apr 88

DATA COLLECTION PROCEDURES

Considerable time and effort were spent initiating, designing, coordinating, and monitoring the LV predictor and training criteria data collection. Numerous briefings were conducted at various points down the chain of command, culminating in several meetings with the POC at each of the eight Reception Battalion sites, several months prior to data collection at that site. From this point until testing began, coordination was taken over by the POC, who was responsible for providing the required troops, space, and necessary equipment. The two primary challenges in preparing each site were (a) fitting 4 hours of testing into an already demanding 72-hour processing schedule, and (b) obtaining adequate space for testing, that met good testing standards, every day for a full year.

A test site manager (TSM) was hired to be in charge of each data collection site, and was supported by from one to as many as eight test administrators. Applications were taken by mail for both positions, and all initial interviewing and hiring was done on site by experienced Project A staff. Detailed test administration manuals were prepared and used as the basis for a one-week training course, conducted at each site for the newly hired personnel. Also, scripts were prepared for administering each test or inventory and test site personnel were trained in their use as well as in handling questions.

Each week the TSM called the Project A staff person in charge of the data collection and reported the number of soldiers tested the prior week, discussed any questions or problems he or she had, and received relevant news or instructions. In addition, each site was required to submit monthly written reports of their testing progress and documentation of any problems that had occurred or events that may have had an impact on test results.

Finally, Project A contractor or ARI staff visited each site from one to three times to monitor the test administration, provide feedback where appropriate, and go over questions or unresolved problems.

SAMPLE SIZES

Predictor Data

The final sample sizes for the Longitudinal Validation predictor data collection are shown in Tables 6-5, 6-6, and 6-7. Table 6-5 shows the number of soldiers at each of the reception battalions who took at least one of the five components of the predictor battery: computer, spatial (paper-and-pencil cognitive), ABLE, AVOICE, and JOB. As the table shows, 49,397 soldiers participated in the administration of the predictor battery. Fort Benning had the largest percentage of the sample, with 28.7 percent, followed by Fort Jackson with 17.6 percent. Forts McClellan, Leonard Wood, and Sill were next with 11.9, 11.5, and 10.3 percent, respectively. Forts Dix and Knox were next, with 8.4 percent and 7.8 percent, respectively, while Fort Bliss had the smallest percentage, 3.7.

177

Table 6-5

Longitudinal Validation: Predictor Data Collected at Each Reception Battalion

Post	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Fort Benning	14,188	28.7	14,188	28.7
Fort Bliss	1,842	3.7	16,030	32.5
Fort Dix	4,160	8.4	20,190	40.9
Fort Jackson	8,700	17.6	28,890	58.5
Fort Knox	3,857	7.8	32,747	66.3
Fort McClellan	5,885	11.9	38,632	78.2
Fort Sill	5,067	10.3	43,699	88.5
Fort Leonard Wood	5,698	11.5	49,397	100.0

Table 6-6

Longitudinal Validation: Predictor Data Collected by MOS

MOS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
11B	14,193	28.7	14,257	28.9
12B	2,118	4.3	16,375	33.1
13B	5,087	10.3	21,462	43.4
16S	800	1.6	22,262	45.1
19E	583	1.2	22,845	46.2
19K	1,849	3.7	24,694	50.0
27E	139	0.3	24,833	50.3
29E	257	0.5	25,090	50.8
31C	1,072	2.2	26,162	53.0
51B	455	0.9	26,617	53.9
54B	967	2.0	27,584	55.8
55B	482	1.0	28,066	56.8
63B	2,241	4.5	30,307	61.4
67N	334	0.7	32,234	65.3
71L	2,140	4.3	34,374	69.6
76Y	2,756	5.6	37,130	75.2
88M	1,593	3.2	31,900	64.6
91A	4,219	8.5	41,349	83.7
94B	3,522	7.1	44,871	90.8
95B	4,206	8.5	49,077	99.4
96B	320	0.6	49,397	100.0
Unk	64	0.1	64	0.1

Longitudinal Validation: Extent of Complete Versus Partial Predictor Data by Reception Battalion and MOS[a]

MOS	Reception Battalion								Total
	Benning	Bliss	Dix	Jackson	Knox	McClellan	Sill	Wood	
11B Complete	4,364		9						4,566
Partial	5,064								5,366
12B Complete			9	1	1			2,995	2,992
Partial				1				25	76
13B Complete		1	4	12	1	2	4,916		4,936
Partial							252		252
14E Complete		674	23	29	6	12		57	781
Partial		18		7				2	10
16S Complete			1	1	678				679
Partial					6				6
19K Complete			7		1,956				1,956
Partial					41				41
27E Complete			6	7	7	8		8	135
Partial						1			1
29E Complete			7	161	12	15		17	212
Partial				44	6	1		45	46
31C Complete			164	722	71	29		38	986
Partial			6	106	6				118
51B Complete			27	13	6	2		398	441
Partial				4				18	18
54E Complete			6	398	6	498		11	961
Partial				71		15			38
63B Complete			44	53	53	218		122	463
Partial			2	11		6		2	25

(Continued)

Table 6-7 *(Continued)*

Longitudinal Validation: Extent of Complete Versus Partial Predictor Data by Reception Battalion[a]

MOS			Reception Battalion							
	Benning	Bliss	Dix	Jackson	Knox	McClellan	Sill	Wood	Total
6B									
Complete			661	842	117	68		464	2,904
Partial			32	188	2			6	117
67B									
Complete			27	1	224	62		14	329
Partial			1	1	3	1			6
71L									
Complete			382	1,424	34	143		2	1,985
Partial			12	228		5			235
71Y									
Complete			221	1,525	188	115		425	2,475
Partial			17	243	2	12			291
88M									
Complete			794	156	28	67		472	1,494
Partial			58	31	1	5		6	96
91A									
Complete		1,129	626	871	444	179		561	3,656
Partial		28	25	213	6	7		6	264
94B									
Complete			616	1,966	211	218		945	3,278
Partial			42	173	2	18		14	243
95B									
Complete			18	2	7	4,070		6	4,186
Partial					1	161			162
96B									
Complete			227	48		2		12	281
Partial			21	17				1	26
Unknown									
Complete		1	18	23	1	11		1	47
Partial		1	2	7				7	17
Total Complete	4,394	1,095	3,354	7,448	3,786	5,726	4,615	5,014	37,434
Partial	37	218	1,256	71	159	252	84	11,963	

[a] Complete predictor data includes data for each soldier on each predictor component: Computer, spatial, ABLE, AVOICE, JOB. Partial predictor data means at least one predictor component is missing for a soldier.

This sample is broken down by MOS in Table 6-6. The 11B is by far the largest MOS in the sample, with 14,193 soldiers representing 28.7 percent of the total. Next is 13B (5,087, 10.3 percent), followed by 91A with 4,219 and 95B with 4,206 (each about 8.5 percent). The four least populous MOS in the sample are: 27E with 139 (0.3 percent), 29E with 257 (0.5 percent), 96B with 320 (0.6 percent), and 67N with 334 (0.7 percent). The MOS for 64 soldiers in the sample remains unknown at this time.

Table 6-7 displays the predictor administration by reception battalion and by MOS, and also provides information on the extent of complete versus partial data. A soldier is counted as "partial" if one or more of the five predictor battery components is missing. For the total sample, 37,434 soldiers (75.8 percent) had complete data -- that is, a record for all five components of the predictor battery for each individual. To accommodate the large number of soldiers being processed at any one time at Fort Benning, the predictor administration was set up to administer the computer component of the battery to only about one-third of the soldiers who came through the reception battalion. If the 9,884 soldiers at Fort Benning who did not take the computer component are excluded, the percentage of soldiers on whom we have complete data increases to 94.7.

End-of-Training Data

The final sample sizes for the Longitudinal Validation end-of-training data collection are shown by post and by MOS in Table 6-8. The number who took the end-of-training measures is shown by whether a soldier took the training achievement test (K3), the rating scales (R), or both (BOTH). Virtually all soldiers took both parts. Table 6-8 shows that 33,863 soldiers out of 34,305 (98.7%) took both end-of-training measures.

Both Predictor and EOT Data

Table 6-9 compares the number of soldiers, by MOS, for whom there are both predictor and end-of-training data. These are the samples that were followed up with the first-tour performance measures. Of the 49,397 soldiers having predictor data, 34,305 (69%) also have end-of-training data. The percentage by MOS ranges from 50 to 92.

Table 6-8

Longitudinal Validation: Extent of End-of-Training (EOT) Data Collected by Post and MOS[a]

BOS	Benning	Bliss	Go	Quarter	Gen Number	Number	Anthon	Bliss	Lee	McClellan	Redstone Arsenal	Roster	BB	Wood	Total
11B K3	277														277
R	18														18
Both	7,702														7,702
12B K3															4
Both							2							1,857	1,857
13C K3													17		17
R													8		8
Both		1											4,264		4,288
16S K3		3													3
Both		878	2												878
19E K3	1							Bliss 1							1
Both								443							443
19K K3	1			1				3					1		3
Both								1,889							1,892
27E K3			1				1				1				1
Both											80				81
29E Both				139											139
31C K3				16											16
Both				656		3	2								652
51B Both											1			349	349
54E K3										2		2			2
Both										656		656			656
55B K3							1				1				1
Both							383	1							384

(Continued)

Table 6-8 (Continued)

Longitudinal Validation: Extent of End-of-Training (EOT) Data Collected by Post and MOS[a]

MOS							Gen. Rsman			Bio.Chem.		Rsman.				Total
63B																
K3			6			4								3	12
R			1												1
Both		487			356								266	1,132
67N																
K3							18	18	10
R							1	1	1
Both							221	221	221
71L																
K3				3		3							3
R													
Both					1,398								1,492
76Y																
K3	18		1	11			11
R	3			2			2
Both	612			1,304		288	1,622
88M																
K3				16		2							21
R				1									3
Both			3,108			1					434	1,390
91A																
K3		3	1	2		3					16
R		1								3
Both		618		947	152		1	3,184
94B																
K3		3		2		1	6		8
R									
Both		3	181	2	3,571		3	3,566
95B																
K3	1	4	16	26	12	33	11	4	65	6	1	3
R											1	1
Both												188
Unknown																
K3	15	3	18	10	19	36		6	8	19	18	2
R	10		5	1	1	2		2	5	1		
Both								475	222	4,068	17	426
Total																
K3	277	3	190	705	36	153		464		4,247		410
R	18		5		1							32
Both	7,816	962	2,585	3,287	4,122	2,944	494	2,031	33,863

[a] K3 = training achievement test only; R = ratings only; Both = both K3 and R.

183

Table 6-9

Longitudinal Validation: Comparison of Soldiers with Predictor Data
Who Also Have End-of-Training Data, by MOS

MOS	Predictor Data			End-of-Training Data				Percent With Predictor and EOT Data[a]
	Complete	Partial	Total	K3	R	Both	Total	
11B	4,308	9,885	14,193	277	18	7,802	8,097	57
12B	2,092	26	2,118	4		1,857	1,861	88
13B	4,835	252	5,087	17	5	4,655	4,677	92
16S	781	19	800	3		578	581	73
19E	578	5	583	1		443	444	76
19K	1,808	41	1,849	3		1,592	1,595	86
27E	138	1	139	1		91	92	66
29E	212	45	257			139	139	54
31C	956	116	1,072	10		652	662	62
51B	441	14	455			349	349	77
54B	881	86	967	2		589	591	61
55B	462	20	482	1		384	385	80
63B	2,094	147	2,241	12	1	1,162	1,175	52
67N	328	6	334	10	1	221	232	69
71L	1,905	235	2,140	3		1,402	1,405	66
76Y	2,475	281	2,756	11	2	1,622	1,635	59
88M	1,494	99	1,593	11	3	1,250	1,264	79
91A	3,935	284	4,219	10	1	3,164	3,175	75
94B	3,279	243	3,522	23	1	1,720	1,744	50
95B	4,101	102	4,203	6		3,580	3,586	85
96B	281	39	320	3		188	191	60
UNK	47	17	64	2		426	428	
Total	37,434	11,963	49,397	410	32	33,863	34,305	69

[a] Computed as total end-of-training data divided by total predictor data.

184

The first-tour LV criterion measures were the same as those used for the Concurrent Validation, except that they were updated as described in this chapter. The 3-year time period between the Concurrent Validation and the Longitudinal Validation raised the issue that some criterion content might be outdated. Equipment and/or procedural changes would require test revisions; changes in MOS responsibilities had the potential of making some tasks obsolete.

Project staff identified relevant changes so that the appropriate revisions could be made. In a few cases where an entire task was obsolete, the task was dropped without replacement. In many cases, revisions were simply a matter of replacing outdated terminology. Updated criterion measures were forwarded to the MOS proponents for a currency review and additional revisions were made on the basis of this review.

Hands-on Measures. Lessons learned from the Concurrent Validation prompted the use of a different format for the hands-on test sheets. An overall effectiveness rating for performance on each task (on a scale of 1 to 7) was added at the end of each task score sheet for hands-on tests in the expectation that it would provide unique task performance information.

After a search for additional first-tour measures that would have relevance for combat readiness, a computer-simulated M16 rifle marksmanship task, the Multipurpose Arcade Combat Simulator (MACS), originally developed for application as a training aid was selected. Using a demilitarized M16 rifle, the soldier "shoots" at targets displayed on a computer monitor. Attached to the barrel of the rifle is a light pen which simulates the path of the rounds and the screen displays a total of 30 targets, some moving and some stationary. Using the MACS, a test of "Engage targets with an M16" was added to the criterion measures for two MOS, 11B and 95B.

Rating Scales. The time period between the two data collections was crucial for MOS 19E (M60 Armor Crewman) because this MOS was being severely scaled back as MOS 19K (M1 Armor Crewman) was being phased in. The two differ with respect to the kind of tank (M60 or M1) that the soldiers operate. To deal with the transition, a job analysis of 19K was conducted and a complete set of criterion measures was developed specifically for this new MOS. The same procedures used for the other MOS (Campbell, 1987b) were followed, with one exception: The 19K MOS-specific rating scales were developed by SMEs from the Armor School and by 19E NCOs. Because of the 19E/K split, the Longitudinal Validation data collection included 10 MOS in Batch A rather than nine.

While there was considerable interest in keeping the Combat Performance Prediction scale, project staff and the Scientific Advisory Group agreed that the version used in the Concurrent Validation was too lengthy. Two alternatives were considered. The first was simply to reduce the number of items in the original summated rating scale of 40 items. The second was to reduce the specific behavioral items to summary dimensions. Three dimensions were derived through empirical and rational analysis, and the new scales were field tested in conjunction with the second-tour criterion measure field tests. Low

reliability estimates for the dimensional ratings led to the decision to retain the original summated scale format, but the total number of items in that summated scale was reduced from 40 to 19.

The final set of Combat Prediction Scale items was selected by considering interrater reliability, internal consistency, and content coverage. That is, items were dropped if their content was covered in another item whose reliabilities were higher, or if their content was specifically technical and therefore covered by another measure, such as a hands-on test or a rating dimension. Three of the original items were deleted because SMEs indicated that the items were not meaningful. The SMEs were field grade officers and senior NCOs with combat or tactical field exercise experience. Another change from the CV version was to use a less cumbersome 7-point scale rather than a 15-point scale.

Personnel File Form. The self-report form for gathering information on administrative records (the Personnel File Form) was updated by reviewing its contents with officers and NCOs who were representatives of the Army's military personnel center. The form was revised to allow soldiers to report administrative actions by pay grade, and to report the date of their last M16 qualification.

Army Job Satisfaction Questionnaire. A new measure developed by the ARI staff was the Army Job Satisfaction Questionnaire. It was intended to provide information that would be potentially useful for predicting attrition and for understanding the relationship of job satisfaction with other constructs investigated. The satisfaction measure was developed in several stages. First, a number of job satisfaction dimensions of relevance to the Army were identified through an extensive search of the literature. Second, items were written to tap each of these dimensions. Items were also written to elicit background information that would help clarify the respondent's frame of reference with respect to his or her perceived satisfaction levels (e.g., reasons for enlisting).

The draft questionnaire was administered to the examinees in the second-tour criterion measure field tests. The Minnesota Satisfaction Questionnaire (MSQ short form) was also administered as a marker instrument. The final set of 18 satisfaction items was selected based on reliability and meaningfulness of the factor structure of the total set. These items assess six aspects of job satisfaction (supervision, co-workers, promotions, pay, work, and Army). Thirteen "frame of reference" items were selected for inclusion on the final questionnaire.

Deleted Measures. Four measures were deleted from the array of Batch A first-tour criterion measures used during the Concurrent Validation. The ratings of performance on the 15 tasks selected for hands-on testing in each MOS were eliminated from the MOS-specific performance rating scales because they were not sufficiently reliable. The common task ratings from the Army-wide rating scales were deleted for the same reason. Two auxiliary measures deleted were the Measurement Method Rating and the Army Work Environment Questionnaire.

Batch Z MOS. With respect to the Batch Z MOS, the school knowledge tests had been submitted to a currency review just prior to the Longitudinal Validation predictor (including training performance) data collection. A

186

second currency review for the criterion data collection was considered neither necessary nor practical. In the currency review, the item pool for each MOS was submitted to the final authority for doctrine on that MOS, the TRADOC proponent, for review and approval. Proponents were free to recommend deletions, additions, and modifications to the test items.

Table 7-1 lists the final array of measures and supplemental information that were administered to and gathered from first-tour examinees during the Longitudinal Validation criterion data collection.

Table 7-1

First-Tour Measures and Supplemental Information Administered to and Gathered From LV Sample

Batch A:	Personnel File Form
	Army-Wide Performance Rating Scales
	MOS-Specific Rating Scales
	Combat Performance Prediction Scale
	Hands-on Tests
	Job Knowledge Tests
Batch Z:	Personnel File Form
	Army-Wide Performance Rating Scales
	Combat Performance Prediction Scale
	School Knowledge Tests
Supplemental Information (Both Batch A and Batch Z):	
	Background Information Form
	Army Job Satisfaction Questionnaire
	Job History Questionnaire
	Physical Requirements Survey*

* Non-Project A measure administered in conjunction with this data collection effort.

Chapter 8
ANALYSIS FOR SECOND-TOUR JOB PERFORMANCE MEASURES

The Project A research plan called for the development of NCO job performance measures which could be used in a second-tour follow-up of two accession cohorts (FY83/84 and FY86/87) for purposes of determining selection/classification/promotion strategies for NCOs. To develop strategies for identifying NCO potential, measures of second-tour job performance are needed. After the criteria are available, the following questions could be examined: To what extent does the Experimental Predictor Battery predict performance beyond the first term of enlistment? Does early performance predict later performance, when additional responsibilities such as supervision and leadership are presumably required? What is the optimal combination of selection/classification test information and first-tour performance data for predicting second-tour performance? How does entry-level training performance relate to later first-tour and second-tour job performance?

Over the life cycle of Project A, the full round of the data collections and analyses necessary to answer these critical questions could not be completed. However, the required job analysis was completed and the criterion development work was begun.

JOB ANALYSIS FOR SECOND TOUR

The specific goals of the job-analytic work were to:

- Describe the major differences between entry-level and higher level performance content, within MOS.

- Describe the major differences across MOS, within higher level jobs.

- Describe the specific nature of the supervisory/leadership component of these higher level jobs.

Once these objectives were achieved, the information would be used to address four questions:

- What should be the content of the new criterion measures?

- What kinds of measurement methods are needed?

- Are separate measures needed for each job? Or are the jobs so similar that the same measures can be applied to all?

- To what extent can measures developed for entry-level soldiers be used among higher level soldiers?

189

The second-tour samples were to be taken from the nine MOS in Batch A and were intended to be subsamples of the FY83/84 and FY86/87 validation samples. The term "second tour" was used by Project A to designate soldiers who have been in the Army between 3 and 5 years. Paygrade will vary from one MOS to another because of differences in density and promotion needs of the Army. Projections indicated that the proportion who would be E5s would be between 20 and 70 percent across MOS. Most others would be E4s; a very few would be E6s.

During FY85, 4,930 soldiers who entered the nine Batch A MOS during the FY83/84 window were tested in the Concurrent Validation sample on the predictor battery, training tests, and first-tour criterion measures. This sample forms the basis of the CVII follow-up. For the second longitudinal follow-up, LVII, the cohort that entered the Army in FY86/87 and the samples that were tested on the Experimental Battery and the training knowledge and performance measures can be followed into their second tour and measured on the job performance criterion measures. These samples are described in Chapter 6 of this report.

SECOND-TOUR JOB ANALYSIS METHODS

By Army policy[4], all soldiers at a higher skill level are responsible for being able to perform all tasks at each lower skill level, as well as the tasks at their current skill level. Consequently, the first-tour job analyses were used as a starting point and additional job analysis information was collected to describe the second-tour changes. In addition, the issue of leadership/supervision performance was of special concern.

To capture both the technical and the supervisory aspects of an MOS, four methods of job analysis were used: task analysis, a standardized questionnaire measure of supervisory and leadership responsibilities, critical incident analysis, and interviews with small groups of senior NCOs.

Task-Based Job Analysis

Specification of the population of second-tour technical tasks proceeded as for first-tour analysis, by combining information from the Soldier's Manuals for each MOS (a Soldier's Manual is prepared by the proponent agency for every skill level within an MOS) and data from the Army Occupational Survey Programs. After being edited for redundancies and level of generality, AOSP items that could not be matched with Soldier's Manual tasks were added to the population of tasks for that MOS. The proponent Army agencies then reviewed the list for completeness and accuracy.

The total task domains for the nine MOS ranged between 153 and 409 tasks each, with an average of 260. To aid in the selection of a representative sample of critical tasks for criterion measurement, judgments of task criticality and performance difficulty were then obtained from 15 officers/

---

[4]Army Regulation 611-201, Enlisted Career Management Fields and Military Occupational Specialties.

190

SMEs who had recent field experience supervising E5s. The officers and SMEs were obtained through the ARI troop support request (TSR) process. The grade, MOS, and experience criteria for the officers and SMEs are laid out in the TSR which is then distributed to the appropriate installation for action. The assigned point-of-contact works with a member of the project staff to iron out the specific details of the data collection, including secondary and tertiary criteria for SME selection.

Also, task clusters were developed for the second tour by using the first-tour clusters as the starting point. That is, the new second-tour tasks were sorted into these same clusters by the project staff. Where no clusters of first-tour tasks were similar to the new second-tour tasks, new clusters were formed.

A Standardized Description of the Supervisory Components of Second-Tour MOS

At the same time that the technical task descriptions were being developed for each MOS, work was also proceeding on a standardized description of supervisory/leadership activities. The item content was derived from two instruments previously developed by ARI researchers: the Supervisory Responsibility Questionnaire, a 34-item instrument based on critical incidents describing effective and ineffective NCO leader behavior (White, Gast, & Rumsey, 1986); and a very comprehensive questionnaire checklist, the Leader Requirements Survey, which contained 450 items and was designed to describe supervisory/leadership activities at all NCO and officer ranks. Both instruments were based on extensive development work and took advantage of the large pool of literature on leader/supervisor behavior (Gast, Campbell, Steinberg, & McGarvey, 1987).

Both questionnaires were administered to NCOs in the nine jobs. Approximately 50 NCOs received the Leader Requirements Survey, and 125 NCOs received the Supervisory Responsibility Questionnaire. All SMEs were asked to indicate the importance of each task for performance at the sergeant (E5) level.

Analysis of the Supervisory Responsibility Questionnaire data confirmed that all the tasks were sufficiently important to be retained. The Leader Requirements Survey importance data were used to select tasks that over half of the respondents indicated were absolutely essential to the sergeant's job, and 53 tasks were retained.

Content analysis of the two task lists resulted in a single list of 46 tasks that incorporated all of the activities on both lists. These tasks, in eight clusters, were added to the second-tour job task list for each of the nine jobs prior to collection of task characteristics data. Later they were made part of the task clustering judgments.

The selection of sample tasks for measurement was based on the importance rating for each task, the performance difficulty and expected performance variability for each task, the frequency of task performance as shown by the AOSP analyses, and the task cluster membership for each task. A

191

Delphi panel of SMEs selected 45 tasks for each job -- 30 technical and 15 supervisory.

The individual panel members first independently selected tasks, using the given targets for each cluster. The choices were tallied and presented to the panel in the second session. They again made independent selections, this time giving reasons for each of their choices. The choices were tallied, the reasons summarized, and the results fed back for consideration through three rounds of independent selections. In the fourth session, the remaining differences were discussed and resolved. Panel members also assigned a complete priority ranking (1-45) for inclusion in the final set.

Critical Incident-Based Job Analysis

To incorporate the Army-wide versus MOS-specific distinction, an inductive critical incident analysis strategy which requires persons familiar with the jobs to generate examples of effective, mid-range, and ineffective performance behavior was again used, as in the first-tour job analyses (Pulakos & Borman, 1986; Toquam, et al., 1986). Content analysis of the examples then yields preliminary dimensions of performance, and an independent retranslation of the examples into the dimensions provides a way of checking on the content validity of the dimension system.

Army-Wide Analysis. Three workshops were conducted in which participants were asked to generate non-MOS-specific examples of what they considered to be specific second-tour performance episodes. A total of 1,000 critical incidents were generated by 172 officers and NCOs. Table 8-1 shows characteristics of the participants in the workshops. These incidents were edited to a common format and then content analyzed to form 12 preliminary dimensions of second-tour Army-wide performance. The nine performance categories that had been developed for the first-tour soldiers were also found in the second-tour analysis; in addition, three generic supervisory dimensions emerged, which suggested that second-tour soldiers do, in fact, perform most of the work that first-tour soldiers perform and also supervise that work. The retranslation results indicated that all 12 of the dimensions resulting from the initial categorization of the incidents should be retained. The second-tour array of 12 Army-wide performance dimensions is shown in Table 8-2.

MOS-Specific Analysis. Development of the second-tour MOS-specific dimensions followed a different procedure and involved a process for revising the existing first-tour MOS-specific rating scales so that they would be appropriate for describing and evaluating second-tour performance.

192

Table 8-1

Participants in Second-Tour Workshops for Generation of Army-Wide
Critical Incidents

Site[a]:		_n_		
	Fort Bragg	102		
	Fort Carson	53		
	Other	3		
	NCOs	_n_	Officers[b]	_n_
Rank:	E-5	19	01	8
	E-6	13	02	26
	E-7	2	03	82
	E-8	1	04	18
			05	2
		_n_		
Gender[c]:	Male	154		
	Female	17		

	Mean	SD
Time in Army:	6.92 years	4.51 years
Time in Supervisory Position:	5.09 years	4.25 years

[a]Fourteen participants left this blank.

[b]One participant left this blank.

[c]One participant left this blank.

193

Table 8-2

Army-Wide Dimensions for Second Tour

A.   Displaying Technical Knowledge/Skill

B.   Displaying Effort, Conscientiousness, and Responsibility

*C.   Organizing, Supervising, Monitoring, and Correcting Subordinates

*D.   Training and Developing

*E.   Showing Consideration and Concern for Subordinates

F.   Following Regulations/Orders and Displaying Proper Respect for Authority

G.   Maintaining Own Equipment

H.   Displaying Honesty and Integrity

I.   Maintaining Proper Physical Fitness

J.   Developing Own Job/Soldiering Skills

K.   Maintaining Proper Military Appearance

L.   Controlling Own Behavior Related to Personal Finances, Drugs/Alcohol, and Aggressive Acts

---

*New leadership/supervisory dimensions for second tour.

To accomplish the revision, a critical incident analysis workshop was conducted with approximately 25 officers and NCOs in each of the nine target jobs (Batch A MOS) to generate examples of effective, average, and ineffective second-tour MOS-specific job performance. Table 8-3 shows characteristics of the participants in the workshops. The number generated for each MOS ranged from 58 to 236 with an average of 180 (Table 8-4). The incidents were then categorized by the project staff, using the first-tour MOS-specific category system as a starting framework. If a second-tour incident did not fit into an existing first-tour category, a new category was introduced. This procedure yielded information regarding what specific category additions or deletions were necessary to describe critical second-tour performance comprehensively.

Almost all of the first-tour MOS-specific performance categories were judged to be appropriate for second-tour MOS. The next step was to examine the content of the incidents to determine whether the performance requirements were appreciably different for second-tour than for first-tour soldiers. If comparisons of the first- and second-tour critical incidents indicated that more was expected of second-tour soldiers than of their first-tour counterparts or that second-tour soldiers were responsible for knowing how to operate and maintain more/different pieces of equipment, such distinctions were incorporated into the second-tour scale anchors.

194

Table 8-3

Participants in Second-Tour Workshops for Generation of MOS-Specific
Critical Incidents, by MOS[a]

		11B	13B	19E	31C	63B	64C[*]	71L	91A/B	95B
Site:	Fort Bragg	11	-	14	1	6	7	8	11	23
	Fort Carson	4	-	4	3	8	4	14	1	15
	Fort Knox	-	-	27	-	-	-	-	-	-
	Fort Hood	-	14	-	-	-	20	-	-	-
	Fort Gordon	-	-	-	17	-	-	-	-	-
	Fort Sam Houston		-	-	-	-	-	-	8	-
	Total	15	14	45	21	14	31	22	20	38

		11B	13B	19E	31C	63B	64C	71L	91A/B	95B
Rank:	NCOs									
	E-4	-	-	2	-	-	-	-	-	-
	E-5	-	-	9	3	-	1	5	-	8
	E-6	-	9	11	12	-	19	1	4	-
	E-7	-	5	5	1	1	5	-	3	-
	E-8	-	-	-	-	-	1	-	1	-
	Officers									
	O1	-	-	-	-	1	-	3	1	3
	O2	-	-	7	1	-	1	4	-	12
	O3	14	-	11	2	12	4	8	8	12
	O4	1	-	-	1	-	-	1	3	3
Gender:	Male	15	14	45	18	13	29	17	18	34
	Female	-	-	-	2	1	2	5	2	4

	11B	13B	19E	31C	63B	64C	71L	91A/B	95B
Mean Time in Army	6.68	11.97	7.74	11.28	6.87	12.06	5.20	7.58	6.40
Mean Time in Supervisory Position	5.17	8.09	5.29	7.31	5.03	7.49	3.80	5.01	4.87

[a]Many of these participants also generated Army-wide critical incidents

[*]MOS 64C subsequently became MOS 88M.

Table 8-4

Second-Tour MOS-Specific Critical-Incident Workshops:
Numbers of Incidents Generated, by MOS[a]

MOS	Number of Participants	Number of Incidents
11B	15	151
13B	14	58
19E	45	236
31C	21	212
63B	14	180
64C[b]	31	184
71L	22	149
91A	20	206
95B	38	234

[a]Many of these participants also generated Army-wide critical incidents.

[b]MOS 64C subsequently became MOS 88M.

For several MOS, the second-tour incidents suggested that MOS-specific supervisory performance categories should be developed. However, in developing categories, care was taken not to duplicate the Army-wide leadership/supervision dimensions and to reflect aspects of supervision that were relevant only to the particular job in question. A total of six MOS-specific supervisory dimensions distributed over five MOS were generated.

For each of the nine MOS, two scale revision workshops were conducted with 10-14 participants (officers and NCOs) in each. Participants considered the validity of the dimension anchors for evaluating second-tour effectiveness, and whether the proposed dimensions were relevant and inclusive of all MOS-specific performance components. Scales were revised if appropriate.

For each MOS a third, or retranslation, workshop was also conducted with approximately 20 officers and NCOs. For 92 percent of the revised incidents, more than 75 percent of the sample categorized them as intended. The dimensions for each Batch A MOS are shown in Table 8-5.

196

Table 8-5

MOS-Specific Dimensions for Second Tour

11B:
Maintaining and Accounting for
  Equipment and Weapons
Supervising Soldiers in the Field
Leading the Team
Navigation
Use of Organic Weapons and
  Equipment
Field Sanitation, Personal Hygiene,
  and Personal Safety
Fighting Positions
Avoiding Enemy Detection
Operating Radio Set
Reconnaissance
Guard and Security Duties
Prisoners of War
Proficiency in Battle

13B:
Loading Out Equipment
Driving and Maintaining Vehicles,
  Howitzers, and Equipment
Transporting, Sorting, Stowing,
  and Preparing Ammunition
Preparing for Occupation/Emplacing
  Howitzer
Setting up Communications
Gunnery
Loading/Unloading Howitzer
Receiving and Relaying Communications
Recording/Record Keeping
Position Improvement

19E:
Maintaining Tank, Tank System, and
  Associated Equipment
Driving and Recovering Tanks
Stowing Ammunition Aboard Tanks
Loading/Unloading Weapons
Maintaining Weapons
Engaging Targets with Tank Weapon
  Systems
Operating Communications Equipment
Preparing Tanks for Field Problems
Assuming Supervisory Responsibilities
  in Absence of Tank Commander

31C:
Inspecting and Servicing
  Equipment
Installing Equipment
Operating Communication
  Devices
Preparing Reports
Maintaining Security
Providing Safe
  Transportation
Preparing for Movement
Managing the RATT Rig

63B:
Inspecting and Testing
  Equipment Problems
Checking Repairs Made
  by Other Mechanics
Troubleshooting
Performing Preventive
  Maintenance Checks and
  Services
Repair
Using/Accounting for Tools
  and Test Equipment
Using Technical References
Equipment Operation
Safety Mindedness
Administrative Duties
Determine Task Requirements
Recovery

71L:
Preparing, Typing, and
  Proofreading Documents
Processing and Distributing
  Documents
Maintaining Office Resources
Establishing and/or Main-
  taining Files IAW MARKS
Correspondence Management
Preparing and Safeguarding
  Classified Materials
Providing Customer Service

(continued)

Table 8-5 (continued)

MOS-Specific Dimensions for Second Tour

88M:
Driving Vehicles
Vehicle Coupling
Checking and Maintaining Vehicles
Using Maps/Following Proper Routes
Loading and Transporting Cargo
Loading and Transporting Personnel
Parking and Securing Vehicles
Performing Administrative Duties
  Recovering Vehicles
Safety-Mindedness
Performing Dispatcher Duties

91A:
Maintaining and Operating Army
Medical Vehicles and Equipment
Maintaining Accountability of
  Medical Supplies and Equipment
Keeping Medical Records
Arranging for Transportation and/or
  Transporting Injured Personnel
Dispensing Medications
Preparing and Maintaining Field
  Site or Clinic Facilities in
  the Field
Providing Routine and Ongoing
  Patient Care
Responding to Emergency Situation
Providing Health Care and Health
  Maintenance Instruction to
  Army Personnel

95B:
Traffic Control and Enforcement
Providing Security
Investigating Crimes and Making
  Apprehensions
Patrolling
Leading the Team in Tactical
  Environment
Promoting Public Image of
  Military Police
Interpersonal Communications
  Skills
Responding to Medical
  Emergencies
Navigation
Avoiding Enemy Detection
Use of Weapons and Other
  Equipment

Job Analysis Interviews

The final job analysis method consisted of short (one-hour) structured interviews that were conducted with small groups (5-8 people) of NCOs in each of the nine jobs. They were asked about the number or percentage of sergeants who would probably be in different duty positions, and about the normal activities of those individuals. They were also asked to indicate how many hours per week those individuals would spend on each of nine supervisory activities and each of two general areas of actual task performance, and how important each of those 11 aspects of the job is for the second-tour NCO.

This information was used primarily to provide information about the relative importance and time spent on leadership/supervision versus technical activities.

RESULTS

Major Differences Between First- and Second-Tour MOS

As defined by the task-based descriptions, the additional second-tour tasks are more difficult and complex, but are of the same general content as the first-tour tasks. The addition of tasks also caused several of the technical clusters to split into more highly differentiated task subgroups.

Another important difference between the first- and second-tour task domains is that MOS-specific leadership clusters were added or expanded in every MOS. In seven of the MOS a new cluster was formed to represent tasks involving either tactical operations leadership or administrative supervision, while in the other two MOS such clusters were greatly expanded due to the addition of new tasks.

As mentioned previously, analysis of the Army-wide critical incidents led to the addition of three dimensions reflecting increased supervisory/leadership responsibilities across all jobs. These three dimensions in effect replaced a single first-tour leadership dimension. All nine of the other Army-wide dimensions that had been developed for first-tour soldiers were replicated for the second-tour job.

Analysis of the MOS-specific critical incidents suggested the retention of all but two of the first-tour dimensions; in three cases, a single first-tour dimension was split into two. Of the 85 first-tour dimensions, 38 (45%) were unchanged. The added technical and supervisory responsibilities for second tour resulted in substantial changes to 44 (52%) of the dimensions, and additional MOS-specific supervisory dimensions were developed for five of the nine MOS. The five MOS-specific supervision/leadership scales are summarized in Table 8-6.

Thus, although the MOS vary in the extent to which supervisor/leadership responsibilities constitute new dimensions of job content, the second-tour soldiers in all MOS are responsible for the performance of their subordinates. The technical content of the jobs is, for the most part, similar to the content of first-tour jobs, although higher proficiency is often expected, and more difficult tasks are frequently added.

199

Table 8-6

Supervisory Performance Categories for Second-Tour
MOS-Specific Scales

	MOS	Performance Category Name
11B	Infantryman	Supervising Soldiers in the Field
		Leading the Team
13B	Cannon Crewman	None
19E	Armor Crewman	Assuming Supervisory Responsibilities in Absence of Tank Commander
31C	Single Channel Radio Operator	Managing the RATT Rig
63B	Light Wheel Vehicle Mechanic	Checking Repairs Made by Other Mechanics
71L	Administrative Specialist	None
88M	Motor Transport Operator	None
91A/B	Medical Specialist/Medical NCO	None
95B	Military Police	Leading the Team in a Tactical Environment

Specific Nature of the Leadership/Supervision Component

As a category of job content, leadership and supervision represent a sizable proportion of the junior NCO position. For example, as judged by the previously described job analysis interview panels, from 35 to 80 percent of the NCO's time is spent on supervisory activities.

Given the substantial nature of the supervision/leadership components, the next step was to attempt a more detailed description of their content in terms of specific dimensions. An item pool was created by first using project staff judgments to identify the tasks in each MOS task domain that represented leadership or supervision content. This total list, summed over the nine Batch A MOS, was edited for obvious redundance and then combined with the 46 items from the Supervisory Responsibilities Questionnaire. This produced a total pool of 341 items (tasks).

200

The pool of 341 individual task items was then content clustered by each of 12 judges selected from the Project A staff. Given the target that the number of content clusters should be between 5 and 15, if possible, each judge sorted the task items into categories and wrote a brief definition for each category (i.e., dimension). Consequently, there were 12 cluster solutions based on individual expert judgment.

Next, the degree of agreement among all 12 judges, in terms of how every pair of items should be clustered, was used as input to an empirical cluster analysis. The results of the cluster analysis were compared to the expert judgment solutions and a synthesized description of specific content dimensions was written by the project staff. To say it another way, a pooled solution was obtained by expert judgment. the results of this pooled solution are shown in Figure 8-1.

---

1. Planning Operations

   Activities that are performed in advance of major operations of a tactical or technical nature. That is, planning for, getting ready for, and developing orders for various kinds of team operations, whether it be combat, support, or technical operations. It is the activity that comes before actual execution out in the field or work place.

2. Directing/Leading Teams

   The tasks in this category are concentrated in the combat and military police MOS. They involve the actual direction and execution of combat and security team activities. They occur out in the field and are heavily dependent on MOS-specific skills. Leading reconnaissance teams, setting up offensive ard defensive positions, carrying out a fire mission, directing the clearing of mine fields, etc. would all be part of this category. They require "real-time" decisionmaking under pressure.

3. Monitoring/Inspecting

   This cluster includes interactions with subordinates that seem to involve keeping an operation going once it has been initiated, such as checking to make sure that everyone is carrying out their duties properly, assisting people to overcome problems, making sure everyone .ias the right equipment; monitoring or evaluating the status of equipment readiness, supply levels, completeness of written reports, adequacy of current operating procedures, etc. This is a non-combat or non-crisis set of activities.

Figure 8-1.   Supervision/Leadership Task Categories Obtained by Synthesizing Expert Solutions and Empirical Cluster Analysis Solution (Page 1 of 2)

4. Individual Leadership

The content of the tasks in this cluster reflects attempts to influence the motivation and goal direction of subordinates by means of goal setting, interpersonal communication, sharing hardships, building trust, etc.

5. Acting as a Model

This dimension is not tied to a specific task content but refers to the NCO modeling the correct performance behavior whether it be technical task performance under adverse conditions, or exhibiting appropriate military bearing. The NCO sets the example.

6. Counseling

A one-on-one interaction with a subordinate during which the NCO provides support, guidance, assistance, and feedback on specific performance or personal problems that the soldiers might be experiencing. It includes counseling on problems of a disciplinary nature.

7. Communication with Subordinates, Peers, and Supervisors

The tasks in this category deal with composing specific types of orders, briefing subordinates on things that are happening, and communicating information up the line to superiors, as well as to peers. Information is disseminated in both written and oral formats.

8. Training Subordinates

A _very_ distinct cluster of tasks that describe the day-to-day role of the NCO as a trainer for individual subordinates. When such tasks are being executed, they are clearly identified as instructional (as distinct from evaluations or disciplinary actions). Involves scheduling, planning, and conducting training.

9. Personnel Administration

This category is made up of "paperwork" or administrative tasks that involve actually doing performance appraisals, making or recommending various personnel actions, keeping and maintaining adequate records, and following standard operating procedures for Army personnel practices.

Figure 8-1. Supervision/Leadership Task Categories Obtained by Synthesizing Expert Solutions and Empirical Cluster Analysis Solution (Page 2 of 2)

Chapter 9
DEVELOPMENT OF SECOND-TOUR JOB PERFORMANCE MEASURES

As described previously, there was considerable job analysis information on which to base second-tour performance measurement. For each MOS, 30 technical (MOS-specific and common) tasks and 15 supervisory tasks were selected to represent the task clusters and all 45 selected tasks were rank ordered in terms of their overall importance to the MOS. The critical incident analysis yielded a portrayal of each MOS in terms of its general and specific critical performance components in both technical performance and leadership, and the series of job analysis interviews yielded a rough estimate of the relative importance and time spent for technical vs. supervisory activities for each MOS. Cluster analyses were used to further explore the specific dimensions of supervisory/leadership performance.

Given available resources, constraints on testing time, guidance from the literature, previous Project A work, and the second-tour job analysis results, a potential set of measurement methods was identified and reviewed by the project staff and the Scientific Advisory Committee. Some of the measurement methods had been used for the first tour and some were newly developed.

As indicated by the second-tour job analyses, there is considerable overlap in job content between first tour and second tour, except that the core technical tasks become more complex and significant components of leadership and supervision are added. Consequently, a number of first-tour measurement methods were modified for second-tour use, and several new measures of supervision and leadership were added.

To accommodate the new supervisory measures, assessment of technical task knowledge and performance (i.e., hands-on and job knowledge tests) was allotted less time than in first-tour performance assessment. Reducing assessment time was judged to be better than eliminating either measurement strategy because (a) highly reliable job knowledge tests can be written for almost any task, and (b) the hands-on tests were designed to have a high degree of content validity. For the job knowledge tests, testing time was reduced by using fewer items for each task. This strategy is not feasible with hands-on tests because the scorable steps within task tests are too interdependent to be selectively eliminated. Consequently, fewer tasks were tested in a hands-on mode relative to the number of tasks so tested for first-tour soldiers.

Three data collections were associated with the development of the second-tour criterion measures. These are outlined in Table 9-1. The table lists the types of individuals involved (i.e., SMEs or job incumbents), testing/workshop locations, and the purpose of each data collection.

Table 9-1

Data Collection Efforts in Second-Tour Criterion Development

Pilot Tests

Location: Proponent Schools

Participants: 4 E6 SMEs and 5 E5 incumbents (per MOS)

Purpose: First tryouts of hands-on tests
               Initial generation of role-play exercises
               Preliminary Situational Judgment Test (SJT) workshops

---

Field Tests

Location: USAREUR, Fort Bragg, Fort Hood

Participants: Primarily second-tour incumbents; 41 to 61 soldiers per MOS

Purpose: Field testing of hands-on tests
               First administration of job knowledge tests
               Administration of experimental version of SJT
               Administration of experimental versions of counseling
                 role-plays
               Development of training role-play
               Administration of draft versions of the second-tour Personnel
                 File Form, the second-tour performance rating scales,
                 the Army Job Satisfaction Questionnaire and marker
                 instrument (MSQ), and two versions of the Combat
                 Performance Prediction rating scale

SJT Workshops

Location: Fort Campbell, Fort Devens, Fort Sam Houston, US Army
                 Sergeants Major Academy (USASMA)

Participants: Senior NCOs (n=56); students and instructors from USASMA
                 (n=91)

Purpose: Generate situations and response alternatives
               Gather effectiveness data on response alternatives
               Review SJT items for realism and appropriateness

---

SECOND-TOUR PERFORMANCE CRITERIA OBTAINED BY
MODIFYING FIRST-TOUR MEASURES

Measures of Technical Task Performance

Because by doctrine[1] Skill Level 2 soldiers (pay grade E-5) are also responsible for Skill Level 1 (covers pay grades E-1 through E-4) tasks, the technical tasks selected for testing first- and second-tour soldiers overlapped to a substantial degree. Development of new job knowledge and hands-on tests for the non-overlapping tasks was modeled after the procedures used for the first-tour tests. The hands-on tests were submitted to pilot testing and a field test before being finalized for administration to the second-tour sample. The first administration of the job knowledge tests took place during the field test data collection.

With respect to the job knowledge tests, item analyses on the field test data were used to identify items which required revision and to reduce the number of items so that the tests could be administered in one hour. Similarly, field test results were used to identify needed revisions to the instructions and scorable steps of the hands-on tests. Also, the field test administration provided the information for determining which hands-on tests were to be administered and which were to be dropped.

Note that the Multipurpose Arcade Combat Simulator that was added to the criterion measure set for first-tour MOS 11B and 95B soldiers was also administered to second-tour soldiers in these MOS.

Rating Scales

As described in the section on the second-tour job analysis, the second-tour Army-wide and MOS-specific performance rating scales were developed using the first-tour scales as a starting point. Information generated through the second-tour job analysis was used to revise these instruments to make them suitable for second-tour soldiers. For example, the Army-wide "NCO potential" scale was replaced with a "senior NCO potential" scale.

Furthermore, a set of scales was added to tap supervisory performance dimensions that were identified in the second-tour job analysis (see Figure 8-1). A list of the areas covered in the rating scales and an example of one of these scales is provided in Figure 9-1.

The Army-wide, MOS-specific, and supervisory performance rating scales were administered during the second-tour field test. No changes to the scales were made as a result of analysis of those data.

A panel of SMEs indicated that the Combat Performance Prediction rating scales as revised for first-tour soldiers would also be applicable for second-tour soldiers. All of the rating scales intended for use with second tour soldiers were administered during the field tests listed in Table 9-1.

---

[1]Army Regulation 611-201, Enlisted Career Management Fields and Military Occupational Specialties.

O ACTING AS A ROLE MODEL

O COMMUNICATION

O PERSONAL COUNSELING

O MONITORING SUBORDINATE PERFORMANCE

O ORGANIZING MISSIONS/OPERATIONS

O PERSONNEL ADMINISTRATION

O PERFORMANCE COUNSELING/CORRECTING

ACTING AS A ROLE MODEL FOR SUBORDINATES

Motivates subordinates to perform effectively through personal example,
including demonstrating high standards of military appearance, bearing, and
courtesy; is a model supervisor for subordinates to look up to by
demonstrating exemplary behavior as a soldier.

Falls below standards and expectations for performance in the category "Acting as a Model" compared to soldiers at same experience level.		Meets standards and expectations for performance in the category "Acting as a Model" compared to soldiers at same experience level.			Exceeds standards and expectations for performance in the category "Acting as a Model" compared to soldiers at same experience level.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)

Figure 9-1. Example of Supervisory/Leadership Performance Ratings.

Personnel File Form II

A Personnel File Form suitable for second-tour soldiers was developed by reviewing the contents of the Personnel File Form for first-tour soldiers with officers and NCOs who were representatives of the Army's Military Personnel Center. In addition to the information gathered on the first-tour version of this instrument, the second-tour form elicits information related to the soldier's promotion and reenlistment background. Three categories were added to the form in an effort to reflect the additional administrative actions appropriate for soldiers in their second tour. These categories were Education, Promotion Boards, and Reenlistment waivers. Army Regulations were reviewed to identify information available on the Promotion Board Worksheet, and officers and NCOs who served on promotion boards were interviewed to answer questions about the NCO promotion process to E-5 and above. A draft version of the second-tour Personnel File Form was administered during the second-tour field test. Only minor changes were made to the form as a result of field test data analyses.

NEW CRITERION MEASURES FOR THE ASSESSMENT OF SECOND-TOUR (NCO) PERFORMANCE

Based on a review of the literature and a careful consideration of the feasibility of additional measurement methods, two new methods were developed for assessing second-tour NCO job performance: role-play exercises and a situational judgment test. The role-play exercises were intended to assess the one-on-one interpersonal skills required for counseling and training subordinates, whereas the Situational Judgment Test (SJT) was intended to cover as broad a range of important supervisory skills as possible within the constraints of a paper- and-pencil format.

Role-Play Exercises

Three role-play simulations were developed:

- Counseling of a subordinate with personal problems.
- Counseling of a subordinate with performance problems.
- Remedial training with a subordinate.

These particular simulations were developed because they cover three of the most critical tasks in the supervisory component in the NCO job, as identified in the job analysis.

The general format for the simulations is for the examinee to play the role of a supervisor. The examinee is prepared for the role with a one-page description of the situation that he or she will be asked to handle. The subordinate is played by a confederate who is trained to act out a detailed role. This confederate also has responsibility for scoring the performance of the supervisor (i.e., examinee).

The information and data for the development of the role-plays came from several sources, including (a) Army NCO training materials, (b) the second-tour pilot tests, and (c) the second-tour field tests. The initial content of the counseling exercises was generated during the first two second-tour

pilot tests. Several promising scenarios were selected for further development.

The initial developmental steps involved the drafting of four documents: (a) a description of the supervisor's role, (b) a short description of the subordinate's role, (c) a set of detailed instructions for playing the part of the subordinate, and (d) a performance rating instrument. Project staff drafted a checklist of behaviors applicable to performance in a counseling situation, to be used as a rating device. This checklist was generated using NCO instructional materials provided by the Army. Participants in subsequent pilot tests tried out the role-plays and provided input for refining them. This was an iterative process with participants in the later pilot tests trying out role-play materials that had already gone through several revisions. These tryouts involved considerable shadow-scoring as a means of evaluating the reliability of the rating checklist.

During the course of the pilot tests, development efforts became focused on two counseling exercises, one in which the subordinate had a personal problem and the other in which the subordinate exhibited a performance-related disciplinary problem. Also during this time the performance checklist evolved into a rating scale format. Anchors for three possible ratings were developed for each performance behavior. The final set of behaviors to be rated underwent considerable refinement.

The first formal tryout for the counseling exercises was during the second-tour field tests. In this setting, the subordinate roles were played by NCOs who were also responsible for hands-on scoring. Each NCO was trained on one of the two counseling exercises. A maximum of one-half day was available for training. During this training, the NCOs learned how to play the roles and how to use the rating scales. During the course of training, NCOs took turns playing the subordinate and supervisor roles. In order to evaluate interrater reliability, at least two raters evaluated each soldier's performance in the simulation exercises. No changes to the role-plays were considered necessary as a result of analysis of the field test data.

The development of the training role-play was somewhat different. The content of the training tasks was determined by having pilot test participants examine the first-tour technical task domains for their MOS and nominate tasks that met the following criteria:

(1) Is relatively complex.

(2) Should allow the trainer to exhibit his or her training skill.

(3) Must have standardized equipment and procedures across locations.

(4) Has minimal performance differences within or across MOS.

(5) Can be trained in 15 to 20 minutes.

(6) Should not be a task that is tested hands-on for second-tour soldiers.

The review indicated that no MOS-specific technical task met all six criteria. The tasks for which minimal MOS differences were expected were too simplistic. For other tasks, large differences in task familiarity were expected both within and across MOS.

Consideration then turned to common soldiering tasks (i.e., first aid, weapons) that might require remedial training. The most likely candidates were associated with drill and ceremony activities. This was a promising area because all soldiers learn drill and ceremonies in basic training, most units perform this function daily, the procedures are the same across posts, and NCOs expressed confidence that this would be an appropriate source of training simulation "tasks". Two drill and ceremony behaviors were selected: the about face and the hand salute. As with the counseling role-plays, materials were prepared to specify the subordinate and supervisor roles in the training exercise and to draft a rating form. Again, the behaviors to be rated were derived from trainer manuals used by the Army. The iterative process of trying out the role-play and revising took place during the field test data collections.

Figures 9-2, 9-3, and 9-4 show the three role-play scenarios, an excerpt from one of the three rating forms, and an outline of the training that would be provided to all subordinate scorers. The plan for administering the role-plays to the second-tour personnel in the CVII sample involved the use of civilians, hired and trained specifically for this data collection, as the role-play confederates. It was decided that the most suitable role-player candidates would be young men with prior military experience. Once hired, role-players were to be given at least 3 days of training in a centralized location.

Prior to administration to the validation sample, the role-play exercise materials were submitted to the U.S. Army Sergeants Major Academy for a proponent review. The reviewers found the exercises to be an appropriate and fair assessment of supervisory skills, and did not request any revisions. At this point, the role-play simulations were deemed ready for administration to the CVII sample.

Situational Judgment Test (SJT)

The purpose of the SJT is to evaluate the effectiveness of judgments about what one should do in typical supervisory problem situations. A critical incident methodology was used to generate situations for inclusion in the SJT, and the SMEs who generated situations and response options were pilot test participants. SMEs were provided with the taxonomy of supervisory/ leadership behaviors generated by the second-tour job descriptions and were given the following criteria for "good" situations:

(1)   It is challenging. Situation should be difficult enough so that not everyone would be likely to know the best response.

(2)   It is realistic.

(3)   There is a best response, or at least some responses are better than others.

209

PERSONAL COUNSELING ROLE-PLAY SCENARIO

Supervisory Problem:

PFC Brown is exhibiting declining job performance and personal appearance. Recently, Brown's wall locker was left unsecured. You have decided to counsel this soldier.

Subordinate Role:

- Soldier is having difficulty adjusting to life in Korea and is experiencing financial problems.

- Reaction to counseling is initially defensive, but will calm down if not threatened. Will not discuss personal problems unless prodded.

DISCIPLINARY COUNSELING ROLE-PLAY SCENARIO

Supervisory Problem:

There is convincing evidence that PFC Smith lied to get out of coming to work today. This soldier has arrived late to work on several occasions and has been counseled for lying in the past. You have instructed Smith to come to your office immediately.

Subordinate Role:

- Soldier's work is generally up to standards, which seems to justify occasional "slacking off." Slept in to nurse a hangover and lied to cover up.

- Initial reaction to counseling is a very polite denial of lying.

- If supervisor insists, soldier admits guilt, then whines for leniency.

TRAINING ROLE-PLAY SCENARIO

Supervisory Problem:

The commander will be observing the unit practice formation in 30 minutes. PVT Martin, although highly motivated, is experiencing problems with the hand salute and about-face.

Subordinate Role:

- Feelings of embarrassment contribute to the soldier's clumsiness.
- Soldier makes very specific mistakes.

Figure 9-2. Supervisory role-play scenarios.

ROLE-PLAY EXERCISES

EXAMPLE OF RATING SCHEME

____ 1. Develops rapport at the start of the session.

         3 =   Opens the interview in a pleasant, nonthreatening manner.

         2 =   Opens the interview in a generally nonthreatening manner but uses a tone of voice or non-verbal actions that leave the subordinate feeling somewhat defensive.

         1 =   Opens the interview in a hostile or threatening manner, leaving the subordinate feeling very defensive from the start.

____ 2. States the purpose of the counseling session clearly and concisely.

         3 =   Outlines all topics to be covered (e.g., the purpose is to discuss the wall locker that was left open last night, any problems the subordinate may be having and what might be done to resolve them, etc.).

         2 =   States at least one general topic to be discussed (e.g., says the purpose is to talk about the subordinate's recent poor performance).

         1 =   Fails to state a purpose for the session; instead, jumps directly into the problems.

Figure 9-3. Example of role-play exercise rating scheme.

A. General briefing and orientation.

B. Distribute supervisor's role, subordinate's role and how to play the subordinate's role. Explain these and have scorers read the materials silently.

C. Summarize the roles. Provide step-by-step instructions about how to play the subordinate's role.

D. Distribute the rating scales, explain the rating system, and have trainees read the scales silently.

E. Review each scale separately, detailing differences between a "3" versus a "2" versus a "1".

F. Break group into pairs and have each pair practice the role-play on their own. The purpose here is to familiarize trainees with the exercise.

G. Bring everyone back together. Select two trainees, one to play the supervisor and the other to play the subordinate. The other trainees observe and score the role play.

H. The group discusses their ratings and resolves discrepancies. Feedback is provided on how well the trainee played the subordinate's role.

I. Steps G and H are repeated until each trainee has had an opportunity to play   the subordinate's role.

J. Break the group into triads. Continue practicing playing the subordinate's role, evaluating the supervisor's performance, and discussing the ratings. Trainer circulates among the groups.

Figure 9-4. Role-player training.

212

(4)   It provides sufficient detail to help the supervisor make a
       choice between possible actions.

(5)   A response to the situation can be communicated in a few
       sentences.

(6)   It relates to the second-tour supervisory duties in any MOS, not
       just one MOS (i.e., it is an Army-wide situation). Some workshop
       participants were also asked to write MOS-specific situations.

Response options were developed through a combination of input from
pilot test SMEs and incumbents at the sergeant level from the field tests.
SMEs wrote short answers (1-3 sentences) to the situations describing what
they would do to respond effectively to each situation. Several strategies
were used to elicit response options, including written alternatives generated
by individuals and alternatives arising out of small group discussions. The
written short answers were content analyzed by research staff and additional
response alternatives generated. Table 9-2 presents the workshops completed
and the work accomplished in generating the initial set of 236 situations.

During the last four workshops, seven to nine E-5 to E-7 SMEs from each
of four MOS scaled the effectiveness levels of 34 responses to 11 situations.
The rationale for generating the preliminary effectiveness scale was to obtain
initial data on possible across-MOS differences in preferred supervisory
style. The grand means of response effectiveness levels differ somewhat by
MOS (Table 9-3), and the correlations between mean MOS ratings (Table 9-4)
show moderately high relationships (Rs = .57 to .73; N = 34).

Additional data were gathered on 180 of the best situations during the
field tests (see Table 9-1). Field test incumbents responded to experimental
items by assessing the effectiveness of each listed response option on a scale
of 1 to 7, and by indicating which option they believed was most and which
least effective. During the analysis of the field test data, the content of
open-ended responses from higher rated versus lower rated soldiers was
compared to help guide the generation of more response alternatives. In
addition, comparisons were made between the perceived effectiveness levels
(i.e., effectiveness ratings) of response alternatives from higher rated
versus those from lower rated soldiers. Response alternatives were revised
and some situations dropped between the first and second field tests. In
addition, the effectiveness level comparisons and response revisions and
situation drops were repeated for the second and third field tests.

Two additional workshops were conducted at Fort Devens and Fort Sam
Houston, with seven to nine NCOs in each. At these workshops effectiveness
scale values were gathered from "expert" NCOs for each response alternative,
the SJT was revised and refined, and a scoring key was developed.

213

Table 9-2

Situation Workshops Completed and Work Acccmplished

Workshop Site	MOS	Situations Generated With "Best" Response	Situations Reviewed By Small Groups	Situations for Which Individual Short Answers Were Written[a]
Fort Campbell	Mixed	74	0	0
Fort Sam Houston	91A/B	40	0	0
Fort Gordon	31C	64	64	0
Fort Sill	13B	71	115	0
Fort McClellan	95B	48	60	0
Fort Ben Harrison	71L	34	25	40
Aberdeen Proving Grounds	63B	32	24	50
Fort Eustis	88M	25	47	40
Fort Benning	11B	35	134	<u>106</u>
Total				236

[a]Seven to nine per situation

Table 9-3

Grand Means of Situation Response Effectiveness by MOS

MOS	Items N	People N	Mean	Standard Deviation
71L	34	7	4.53	1.22
63B	34	8	4.64	1.42
83M	34	7	4.76	1.46
11B	34	9	5.42	1.12
Total Sample	34	31	4.89	1.13

Table 9-4

Intercorrelations of Vectors of Item Means for Each MOS
and for the Total Sample (N = 34)

	Total Sample	71L	63B	88M	11B
Total Sample	1.00				
MOS 71L	.83	1.00			
MOS 63B	.91	.70	1.00		
MOS 88M	.83	.57	.71	1.00	
MOS 11B	.87	.67	.73	.60	1.00

A final set of 35 test items was selected on the basis of four criteria:
(a) good agreement among SMEs on "correct" responses, less agreement among
incumbents; (b) item content representation; (c) good distractors; and (d)
USASMA proponent feedback. There are three to five response options per item.
The instructions and an example item are shown in Figure 9-5. Examinees are
asked to indicate the most and least effective response alternative to each
situation. The Reading Grade Level of the test, as assessed using the FOG
index, is seventh grade. Subsequent to Project A, various scoring schemes
will be developed using the effectiveness ratings for response alternatives
obtained in the field tests and the item analyses to be conducted using CVII
data. These scoring approaches include weighting an examinee's "most effec-
tive" choice for a situation by that response alternative's effectiveness
scale values (provided by SMEs).

In addition to providing SMEs to generate scaling data, USASMA provided
a proponent review of the final test. As with the role-play exercises, USASMA
reviewers considered the SJT to be a fair and appropriate method for assessing
supervisory performance. The SJT also shares with the role-plays the limita-
tion that it was not thoroughly field tested prior to administration to the
CVII sample. Consequently, the CVII data collection is most appropriately
considered a field test.

In this booklet, you will be presented with a series of supervisory situations. These are situations in which a first-line supervisor might find him/herself. After each situation several possible responses to that situation are listed.

Read each situation and the responses listed. Then decide which of these possible responses would be the most effective. Place an "M" in the box next to the most effective response.

Next decide which of these possible responses is the least effective. Place an "L" in the box next to the least effective response. The boxes in front of the remaining response alternatives should be left blank.

Below is an example of an item which has been completed properly.

> You are a squad leader. Over the past several months you have noticed that one of the other squad leaders in your platoon hasn't been conducting his CTT training correctly. Although this hasn't seemed to affect the platoon yet, it looks like the platoon's marks for CTT will go down if he continues to conduct CTT training incorrectly. What should you do?

L   a.   Do nothing since performance hasn't yet been affected.

       b.   Have a squad leader meeting and tell the squad leader who has been conducting training improperly that you have noticed some problems with the way he is training his troops.

       c.   Tell your platoon sergeant about the problem.

M   d.   Privately pull the squad leader aside, inform him of the problem, and offer to work with him if he doesn't know the proper CTT training procedure.

You may not agree with the placement of the "M" and the "L" for this item, but this example shows you how these items should be completed.

In summary, for each item you will place an "M" for Most effective next to one response alternative, and an "L" for Least effective next to another response alternative. The boxes in front of the rest of the response alternatives will be left blank. Please use only one "M" and only one "L" per item.

Figure 9-5. Situational Judgment Test Instructions.

SUPPLEMENTAL INFORMATION

Several instruments designed to obtain supplemental information were included in the set of second-tour measures:

Army Job Satisfaction Questionnaire. The Army Job Satisfaction Questionnaire was administered to both first-tour and second-tour soldiers.

Job History Questionnaire. A Job History Questionnaire was included in the final set of second-tour criterion measures. This instrument is the same as that used for first-tour soldiers except that it lists the tasks selected for second-tour soldier testing.

Background Information Form. As with the first-tour soldiers, it was necessary to gather a few items of descriptive information on each examinee (e.g., Social Security Number). The Background Information Form developed for second-tour soldiers also included several questions related to the extent of the examinee's supervisory experience.

Measurement Method Rating. Because two novel testing strategies were to be incorporated into the set of second-tour criterion measures, a Measurement Method Rating form was also included. This form is similar to the one used during the Concurrent Validation, but was modified to reflect the new testing methods.

A list of the complete array of second-tour measures and supplemental information is provided in Table 9-5.

Table 9-5

Second-Tour Criterion Measures and Supplemental Information

Criterion Measures:

    Personnel File Form II[a]
    Army-Wide Performance Rating Scales II
    MOS-Specific Rating Scales II
    Combat Performance Prediction Scales
    Supervisory Simulation Exercises
    Situational Judgment Test
    Hands-on Tests II
    Job Knowledge Tests II

Supplemental Information:

    Background Information Form II
    Army Job Satisfaction Questionnaire
    Job History Questionnaire II
    Measurement Method Rating II

---

[a] "II" indicates that this version is specific for second-tour soldiers.

Chapter 10
LONGITUDINAL VALIDATION CRITERION DATA COLLECTION

The longitudinal criterion data collection began in July 1988 and was completed in February 1989. The primary purpose of the data collection was to test first-tour soldiers who had taken the Experimental Predictor Battery as they entered the Army (the "LVI" sample). A second purpose of the data collection was to collect second-tour performance data (the "CVII" sample) from soldiers who had also participated in the Concurrent Validation (the "CVI" sample). As with the Concurrent Validation, data collections were planned for 13 CONUS installations and USAREUR. The data collection schedule at those installations is shown at Table 10-1.

Table 10-1

LVI/CVII Data Collection Test Dates, 1988-89

Post	Dates		
Fort Lewis	11 Jul- 5 Aug		
Fort Bragg	18 Jul-17 Aug		
Fort Riley	19 Jul-11 Aug		
Fort Hood	25 Jul-24 Aug		
Fort Ord	6 Sep-30 Sep		
Fort Bliss*	15 Sep-29 Sep and	9 Jan-20 Jan	
Fort Campbell	3 Oct-28 Oct		
USAREUR	10 Oct-16 Feb		
Fort Knox*	11 Oct-23 Nov		
Fort Sill*	17 Oct-28 Oct		
Fort Polk	17 Oct-10 Nov		
Fort Benning*	14 Nov-18 Nov and	5 Dec-9 Dec and	19 Dec-20 Dec
Fort Carson	2 Dec-16 Dec		
Fort Stewart	3 Jan- 3 Feb		

* Indicates first tour only

DATA COLLECTION PROCEDURES

Advance Coordination

Advance site coordination for each military installation was accomplished via extensive correspondence (written and phone) and either one or two test site visits. The first site visit provided briefings to post commanders and/or their representatives to clarify the data collection objectives, activities, and requirements. One to two weeks prior to the actual data collection, project staff members visited the installation to examine the test site and discuss equipment, supplies, and other special requirements for the data collection and set-up of the hands-on test stations.

Preceding Page Blank

Using updated listings from the Army's Worldwide Locator Service, post POCs were given a list of the names of target examinees who were shown to be stationed on that post. The POCs used this list to identify the soldiers whom they needed to schedule for testing. To ensure that sufficient data from each MOS were collected, the samples were augmented with additional soldiers who were not in the original sample, but were in the appropriate MOS with the requisite time in service to make them comparable to the characteristics of the target examinees. The operational definition for first- and second-tour soldiers for this data collection was: First-tour soldiers entered the service between 20 Aug 86 and 20 Nov 87; second-tour soldiers entered the service during the period 1 Jul 83 to 30 Jun 84.

Test Site Staffing and Training

Generally, each test site required the following personnel:

Test Site Manager (TSM)	1
Hands-on Managers (HOM)	2
Hands-on Assistants	2
Paper-and-Pencil, Rating Scale, and Role-Play Administrators	5

Additionally, the Army posts provided eight NCOs per MOS to administer and score hands-on tests.

Training of Primary Staff. Most of the nonmilitary test site staff were permanent employees of the contractor consortium. However, a significant number of additional primary staff had to be hired on a temporary basis because of the special requirements imposed by the role-plays. These additional test site personnel played the roles of problem subordinates in the role-play simulations and served as the role-play scorers. Much of the training for in-house staff members took place during the Concurrent Validation and the second-tour field tests. In addition, a formal training program was conducted just prior to the start of the LVI/CVII data collection trips. In preparation for the formal training program, three manuals were constructed: (a) a Test Administrator's Manual, (b) a Test Site Manager's Manual, and (c) a Hands-On Manager's Manual. The instructional materials included the following elements:

- Project A background
- Things to know on an Army post (e.g., rank insignia)
- Criterion measure administration (including dry runs)
- Maintaining integrity of tests and data

The training materials were covered in a 2-day training session. The individuals who were designated role players had an additional 3 days of intensive role-play actor/scorer instruction (see Figure 9-4).

The individuals selected for TSMs and HOMs were generally more experienced than the other test site members. The HOMs, particularly, had to be familiar with the equipment and procedures involved with the tests they would administer for each MOS. For some MOS, such familiarity takes a significant amount of experience to acquire because of factors such as

complexity or diversity of equipment that is used. A good example is Light Vehicle Repairer (MOS 63B). An HOM for this MOS must be familiar with many different vehicles so that when the requested vehicle for a task test is unavailable (as will invariably happen from time to time), he or she can specify a suitable alternative.

Hands-On Scorer Training. Training of all military scorers at the test sites was conducted in conjunction with the actual data collection. NCO scorers for each MOS received from 1 to 2 days of hands-on test administration training prior to the test administration (one day for first-tour tests plus one day for second-tour tests, if applicable). This training was provided on an MOS-specific basis by the HOM for that MOS.

The training followed the procedures that had been developed for the CV data collection (Campbell, 1985). This program is designed not to train the NCOs in how to perform the tasks, but to ensure that each NCO scorer has a fairly high degree of scoring expertise and familiarity with the task tests.

Daily Logistics

The schedule for administering the criterion measures was arranged so that no more than two Batch A MOS (first- and/or second-tour) would be assessed on a given day. Batch Z testing was usually conducted on days when NCO scorers were being trained to administer HO tests to the Batch A examinees. The general plans for administering the criterion measures to these three groups of examinees (Batch A first tour, Batch Z first tour, Batch A second tour) are outlined below. Batch A testing required one day per examinee and Batch Z testing required one-half day per examinee.

All test administration sessions began in the same way. The examinees assembled and roll was taken so that a search could start for any missing personnel. A project staff member would then introduce the soldiers to Project A and review the activities in which they would participate throughout the day. The Privacy Act was read aloud to the soldiers at this time. Soldiers also identified those individuals for whom they would be able to provide peer ratings. If there were 20 or more soldiers in a Batch A MOS or if there were both first- and second-tour examinees present, the total group was divided appropriately into subgroups.

Batch A First Tour. The Batch A first-tour assessment schedule is shown in Figure 10-1. The HO testing was set up to process a maximum of 20 soldiers in a 4-hour period. Eight NCO scorers were needed to meet this schedule. Thus, when there were more than 20 first-tour soldiers from a given MOS to be tested, they were divided into two groups. One group took the HO tests in the morning while the other group took the other criterion measures. After lunch, roll was taken again and the activities of the two groups were reversed. The HO tests for the two MOS were administered in separate locations; however, the written tests and ratings were often administered to both MOS together. This minimized requirements for test site staff personnel.

221

Time	MOS A			MOS B	
	A-1	A-2		B-1	B-2
0730	In-Processing			In-Processing	
0800	HO	JK		HO	JK
0900	HO	JK		HO	JK
1000	HO	X1		HO	X1
1100	HO	X1		HO	X1
1200	Lunch			Lunch	
1300	JK	HO		JK	HO
1400	JK	HO		JK	HO
1500	X1	HO		X1	HO
1600	X1	HO		X1	HO

Legend: HO = Hands-on Tests
JK = Job Knowledge Tests
X1 = Personnel File Information Form
Job History Questionnaire
Peer Ratings (AN/MOS-specific BARS & Combat Scales)
Physical Requirements Survey

NOTE: This schedule assumes four groups of examinees (maximum n=20); two groups for each of two MOS.

Figure 10-1. Batch A first-tour criterion administration schedule.

Batch A Second Tour. On days when second-tour soldiers were being tested, there was normally one group of first-tour soldiers and one group of second-tour soldiers per MOS. The general test administration plan that was used when second-tour examinees were involved is shown in Figure 10-2. The second-tour schedule differs from the first-tour schedule in that one-half of the day was devoted to a combination of 3 hours of HO testing and 1 hour of supervisory simulation exercises, and the other one-half day was devoted to a somewhat different combination of written tests and ratings. Specifically, the time devoted to the job knowledge test was reduced from 2 hours to 1 hour to make time for the 1-hour Situational Judgment Test.

Time	1st Tour MOS A	2nd Tour MOS A	1st Tour MOS B	2nd Tour MOS B
0730	In-Processing		In-Processing	
0800	HO	JK	JK	HO
0900	HO	X2	JK	HO
1000	HO	X2	X1	HO
1100	HO	S	X1	HO
1200	Lunch		Lunch	
1300	JK	HO	HO	JK
1400	JK	HO	HO	X2
1500	X1	HO	HO	X2
1600	X1	HOM	HO	SM

Legend: HO - Hands-on Tests
JK - Job Knowledge Tests
S - Situational Judgment Test
X1 - Personnel File Information Form
Job History Questionnaire
Job Satisfaction Questionnaire
Peer Ratings (AW/MOS-specific BARS & Combat Scales)
Physical Requirements Survey
X2 - Personnel File Information Form
Job History Questionnaire
Job Satisfaction Questionnaire
Peer Ratings (AW/MOS-specific BARS & Combat Scales) or
ABLE
M - Measurement Method Ratings

NOTE: This schedule assumes four groups of examinees (maximum n=20); two
groups (one first tour, one second tour) for each of two MOS.

Figure 10-2. Batch A first-/second-tour criterion administration schedule.

There was an expectation that a significant percentage of second-tour
soldiers would not be able to provide peer ratings. One of the primary
problems is that soldiers at this level often work much more autonomously than
their first-tour counterparts; another problem is that second-tour soldiers
were tested in very small groups, thus decreasing the likelihood that there
were many pairs of co-workers. Plans were therefore made to make the most of
the time that examinees not making peer ratings would have available. The
Project A biodata predictor, ABLE, was selected as the instrument examinees
would complete if they could not make peer ratings. This instrument was
chosen because (a) many of the second-tour examinees would be supplemental

223

(i.e., no Project A predictor data would be available for them), and (b) the Army's decision to implement ABLE made it a prime candidate for additional data collection.

Batch Z. The maximum number of Batch Z soldiers who were tested at one time was generally 30. The test administration schedule appears in Figure 10-3.

Session		
Morning	Afternoon	Activity
0730	1230	In-Processing
0800	1300	School Knowledge Tests
0900	1400	School Knowledge Tests
1000	1500	PSD
1100	1400	R

Legend: P - Personnel File Information Form
      S - Job Satisfaction Questionnaires
      R - Peer Ratings (AW BARS & Combat Scales)
      D - Physical Requirements Survey

NOTE: This schedule assumes four groups of examinees (maximum n=20); two groups for each of two MOS.

Figure 10-3. Batch Z criterion administration schedule.

Supervisor Ratings. The goal was to obtain two supervisor ratings for each examinee. Supervisor raters were identified with the assistance of the examinees and the NCO support staff. One of the project staff was responsible for coordinating efforts to (a) identify the supervisors, (b) schedule rating administration sessions with them, and (c) administer the supervisory rating sessions. The supervisory rating sessions ran concurrently with the other data collection and scorer training activities. Supervisors were requested to report on the same day as their subordinates.

Assessment of Interscorer Agreement (Hands-on and Role-Play)

Although some effort was devoted to assessing hands-on test reliability in early Project A data collection efforts, the information was inadequate for providing a reasonable assessment of the interrater reliability of these measures. Consequently, shadow-scoring efforts were incorporated into the LVI/CVII data collection. Interrater reliability estimation efforts focused on the first-tour HO tests for two Batch A MOS (11B and 91A). Collecting shadow-scoring data for these two MOS was arranged at several data collection sites and required a total of 12 scorers (instead of the formally requested eight) for each of these MOS. All scorers were trained to run two

224

of the eight HO testing stations. Four extra scorers were designated as shadow-scorers, and they followed a randomly selected subset of examinees from station to station. Thus, for a subset of 11B and 91A examinees, performance on all of their HO tests was rated by two scorers.

Shadow-scoring data for the supervisory simulations were also collected at test locations in USAREUR. This was possible because there were always at least four trained role-players at each of these test sites and only three simulations were being conducted at any one time. Thus, one individual was available to observe one of the ongoing simulations and provide an independent set of scores for the examinee. Again, the issue was whether the performance ratings assigned by the role-player scorers are reliable across different scorers. Pending data entry, the sample size and analysis results are not known.

SAMPLE SIZES

Pending data entry, exact sample sizes are unknown. Table 10-2, however, provides reasonable estimates of the LVI/CVII sample sizes. The figures are broken down by installation, MOS, and tour (first or second).

Table 10-2

Project A LVI/CVII Estimated Data Collection Totals

				FIRST-TOUR SOLDIERS: Batch A						
Post	11B	13B	19E	19K	31C	63B	71L	88M	91A	95B	Total
Lewis	73	65	27	-	29	94	62	31	87	45	513
Riley	-	38	-	57	33	40	27	70	46	56	317
Bragg	145	119	-	-	87	81	92	41	84	-	649
Hood	-	72	3	299	74	85	114	73	67	-	787
Ord	139	39	-	-	20	20	14	29	51	9	321
Bliss	-	13	-	44	25	29	15	33	35	-	194
Campbell	154	84	-	-	44	73	41	67	67	41	571
USAREUR	181	184	-	243	112	143	138	173	137	164	1475
Sill	-	157	-	-	27	31	39	34	15	12	315
Knox	-	-	48	41	-	22	18	18	29	21	197
Polk	53	31	1	103	22	44	26	42	54	31	407
Benning	51	12	21	5	4	7	31	31	48	30	240
Carson	48	53	153	-	38	39	40	38	70	17	496
Stewart	62	49	-	28	17	45	20	52	38	27	338
Total	906	916	253	820	532	753	677	682	828	453	6820

(Continued)

Table 10-2 (Continued)

Project A LVI/CVII Data Collection Totals

					SECOND-TOUR SOLDIERS:		BATCH A			
Post	11H	13B	19F	19K	31C	63B	71L	88M	91A	95B	Total
Lewis	19	14	8	-	12	17	17	17	14	9	127
Riley	-	14	-	-	5	11	7	14	5	16	72
Bragg	13	18	-	-	11	9	11	13	11	-	86
Hood	-	13	-	-	15	11	12	14	8	-	73
Ord	9	9	-	-	5	8	7	6	6	7	57
Campbell	21	18	-	..	12	10	9	15	15	10	110
USAREUR	28	32	-	-	31	19	38	52	28	56	284
Polk	15	13	-	10	5	13	7	13	6	15	97
Carson	18	16	25	-	-	11	-	-	-	16	86
Stewart	4	15	-	-	7	7	4	-	12	12	61
Total	127	162	33	10	103	116	112	144	105	141	1053

BATCH Z

Post	12B	16S	27E	29E	51B	54B	55B	67N	76Y	94B	96B	Total
Lewis	47	32	-	13	24	34	31	17	78	77	12	365
Riley	47	28	5	2	23	15	22	-	23	51	11	227
Bragg	89	42	3	10	9	46	24	12	94	93	9	431
Hood	48	62	5	9	15	61	38	36	81	78	20	453
Ord	36	12	3	-	-	18	5	6	27	39	2	148
Bliss	14	7	-	-	-	16	-	4	30	30	-	101
Campbell	109	27	4	7	-	35	15	14	72	70	17	370
USAREUR	190	162	52	54	73	170	83	53	124	155	26	1142
Sill	50	-	-	-	-	3	-	-	45	29	-	127
Knox	29	-	-	-	-	7	-	-	44	15	-	95
Polk	88	45	8	4	13	30	-	20	42	45	7	302
Carson	47	24	3	6	10	32	10	14	48	63	9	266
Benning	12	9	-	-	27	8	30	6	38	28	4	172
Stewart	34	22	7	7	8	23	21	15	43	58	11	249
Total	840	472	90	112	212	498	279	197	789	831	128	4448

Batch A First-Tour total	6,820
Batch Z Total	4,448
Total First Tour	11,268
Total Second Tour	1,053
Grand Total	12,311

Chapter 11
EPILOGUE

This final statement on Project A work begins with a brief history of selection and classification, to characterize the context and sequences in which the Project A research has been performed. The chapter closes with a summary list of Project A products and results, in terms of both scientific achievement and practical application.

A BRIEF HISTORY OF SELECTION AND CLASSIFICATION

Formal personnel selection and classification using standardized measures of individual differences actually began in 1115 B.C. with the system of competitive examinations that led to appointment to the bureaucracy of Imperial China (DuBois, 1964). It soon included the selection/ classification of individuals for particular military specialties, as in the selection of spear throwers with standardized measures of long-distance visual acuity (e.g., identification of stars in the night sky).

Systematic attempts to deal with selection/classification issues have been a part of military management ever since. Military organizations are virtually unique in their need to make large numbers of complex personnel decisions in a short space of time. However, the centrality of criterion-related validation to a technology of selection and classification was not fully articulated until World War II, and research and development sponsored by the military has been the mainstay of growth in that technology from then to the present.

The contributions of military psychologists during World War II are well-known and well-documented. The early work of the Personnel Research Branch of The Adjutant General's Office was summarized in a series of articles in the _Psychological Bulletin_ (Staff, PRB, AGO, 1943 a, b, c, d, e, and f). Later work was published in Technical Bulletins and in such journals as _Psychometrika, Personnel Psychology,_ and _Journal of Applied Psychology_. The Aviation Psychology Program of the Army Air Forces issued 19 volumes, with a summary of the overall program presented in Volume I (Flanagan, 1948). In the Navy, personnel research played a smaller and less centralized role, but here too useful work was done by the Bureau of Naval Personnel (Stuit, 1947).

Much new ground was broken. There were important advances in the development and analysis of criterion measures; Thorndike's textbook based on his Air Force experience presented a state-of-the-art classification and analysis of potential criteria (Thorndike, 1949). Improvements were made in rating scales. Forced-choice methods were developed by the Personnel Research Branch; checklists based on critical incidents were used in the AAF program. The sequential aspect of prediction was articulated and examined; tests "validated" against training measures (usually pass/fail) were checked against measures of success in combat (usually ratings or awards). At least one "pure" validity study was accomplished, when the Air Force sent 1,000 cadets into pilot training without regard to their pilot stanine derived from the classification battery. This remains one of the few studies that could report validities without correcting for restriction of range. Historically, 1940 to 1946 was a period of concentrated development of selection and classification

227

procedures, and the further accomplishments of the next several decades flowed directly from it.

In part, this continuity is attributable to the well-known fact that many of the psychologists who had worked in the military research establishments during the war became leaders in the civilian research community after the war. In part, it is attributable to the less widely recognized fact that the bulk of the work continued to be funded by military agencies. The Office of Naval Research, the Army's Personnel Research Branch (and its successors), and the Air Force Human Resources Research (HRR) installations were the principal sponsors.

The bibliography is very long. Of special relevance to the present project is the pioneering work on differential prediction by Brogden (1946a, 1951) and Horst (1954, 1955); on utility conceptions of validity by Brogden (1946b) and Brogden and Taylor (1950); on the "structure of intellect" by Guilford (1957); on the establishment of critical job requirements by Flanagan and associates (Flanagan, 1954); and on the decision-theoretic formulations of selection and classification developed by Cronbach and Gleser (1957) for the Office of Naval Research. The last of these (Psychological Tests and Personnel Decisions) was hailed quite appropriately as a breakthrough--a "new look" in selection and classification--but the authors were the first to acknowledge the relevance of the work of Brogden and Horst cited above. It was the culmination of a lengthy sequence of development.

Project A was carried out in the context of this impressive history, and it has become another milestone. It is by far the most comprehensive personnel research and development project ever attempted. It is unique in that a complete personnel system is being examined at one time. The jobs (MOS) to be studied were sampled representatively from the complete population, now predictor measures were sampled systematically from the complete domain of potential information, and job performance was assessed as thoroughly as possible with multiple measures. Given this data base, and using state-of-the-art analytic techniques, the functioning of the complete selection/classification decision process can be modeled and actually evaluated under various goals or constraints. Project A is truly a landmark in personnel research.

PROJECT A PRODUCTS AND RESULTS

The Project A products in the following list are of two general kinds--products for the "science" (personnel research) and products for the organization (the Army). The list is intended to move from the scientific to the applied. However, the distinction is not always easy to make since many products are useful for both.

(1)   There exist, in technical report form, comprehensive reviews of all validity evidence pertaining to selection and classification for skilled jobs. These are the most comprehensive such reviews ever done.

(2)   The question of whether the Armed Services Vocational Aptitude Battery (ASVAB) does or does not predict job performance (in addition to training performance) has been answered definitively, in the affirmative. The Army and the Department of Defense are now in a firmer position to support their quality goals. In

228

addition, it is now known what aspects of performance ASVAB predicts best and which aspects of performance could be predicted better with other types of selection instruments.

(3) A set of new experimental tests has been developed to measure non-cognitive, psychomotor, perceptual, and cognitive characteristics that are not now measured by the ASVAB. The scope of Project A made it possible to examine virtually the entire domain of selection information, sample from it, and investigate the basic incremental validity produced by each major piece of information.

(4) Using much more comprehensive samples than ever before, new ASVAB Aptitude Area composites have been developed which are firmly data based and empirically defensible.

(5) The results of an expert judgment study of expected correlations between predictor constructs and performance factors are available. In brief, a large sample of personnel experts considered the population of predictor and criterion variables appropriate for entry-level jobs and forecasted what the validity coefficients would be. The consistency in the judgments and their correspondence with known data points make these a potentially valuable tool for future test selection and synthetic validation work.

(6) Much has been learned about the nature of performance in entry-level skilled jobs (e.g., first-tour MOS). We now have a much clearer idea of what major factors constitute performance and how they can be measured. The "criterion problem" is better understood. This knowledge base should better inform future enlistment and promotion policy, as well as future personnel research.

(7) The Concurrent Validation data support the assertion that supervisor ratings of subordinate performance have considerable construct validity if a careful measurement procedure is followed. The data also support the conclusion that supervisors seem to assess both the technical performance of individuals and their general dependability/motivation at the same time.

(8) Within the limits of the Concurrent Validation design, the incremental validity of appropriate ABLE scales for predicting the "will do" components of performance has been demonstrated.

(9) The potential of the AVOICE for differentially predicting "can do" performance in combat vs. technical vs. administrative support MOS has been established. What is needed to make this finding operational is empirical scoring keys.

(10) The Project A job/task analysis procedures worked well and can be used by the Army in the future to develop training curricula, Skill Qualification Test (SQT) content, performance measures, and field exercises. The job analysis summaries for each MOS serve as a model for future job analysis work in the Army as well as in the public and private sectors.

229

(11) Advanced Individual Training (AIT) achievement measures have been developed for 21 MOS. The training measures will allow a determination of whether training performance predicts job performance, and whether it does so differentially for different groups of trainees (race, gender), and different groups of MOS (combat, combat support, combat service support).

(12) The package of rating scale administration procedures can be used in future personnel research in the Army. A major effort in the Project A research was to develop an effective and efficient set of procedures for administering performance rating scales to large numbers of people. These procedures and the package of materials can be adapted for use in other Army personnel research where ratings of many persons are required.

(13) The Supervisory Description Questionnaire (which came out of second-tour job analyses work) is a useful instrument for future work in the design of leadership training or the evaluation of leadership/supervisor performance. The questionnaire is based on a clear rationale and is straightforward to use.

(14) Project A developed a common utility scale for making comparisons across MOS and performance levels within MOS. Although it does not speak to marginal utility issues, it can be used to enhance the comparison of alternative selection/classification procedures.

(15) One very real, and very important product, is the Project A data base itself. It is by orders of magnitude the largest and most completely documented personnel research data base in existence.

References

Borman, W. C. (1986). Behavior-based rating scales. In R. A. Berk (ed.), Performance assessment methods and applications, p. 75. Baltimore: Johns Hopkins Press.

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1986). Development of a model of soldier effectiveness (ARI Technical Report 741). (AD A191 625)

Bownas, D. A., & Heckman, R. W. (1976). Job analysis of the entry-level firefighter position. Minneapolis, MN: Personnel Decisions.

Brogden, H. E. (1946a). An approach to the problem of differential prediction. Psychometrika, 11, 139-154.

Brogden, H. E. (1946b). On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 37, 65-76

Brogden, H. E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educational and Psychological Measurement, 11, 173-195.

Brogden, H. E., & Taylor, E. K. (1950). The dollar criterion--applying the cost accounting concept to criterion construction. Personnel Psychology, 3, 133-154.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). (AD A182 645)

Campbell, J. P. (ed.). (1987a). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1985 fiscal year (ARI Technical Report 746). (AD A193 343)

Campbell, J. P. (ed.). (1987b). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1986 fiscal year (ARI Technical Report 792). (AD A198 856)

Campbell, J. P. (ed.). (1989a). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1987 fiscal year (ARI Technical Report 862). (AD A219 046)

Campbell, J. P. (ed.). (1989b). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1988 fiscal year (ARI Research Note 91-34). (AD A233 750)

Campbell, J. P., Dunnette, M. D., Arvey, R., & Hellervik, L. (1973). The development and evaluation of behaviorally based rating scales. _Journal of Applied Psychology, 57_, 15-22.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. _Applied Psychological Measurement, 2, 4_, 595-601.

Cronbach, L., & Gleser, G. (1957). Psychological tests and personnel decisions. University of Illinois Press.

Davis, R. H., Davis, G. A., Joyner, J. N., & deVera, M. V. (1986). Development and field test of job-relevant knowledge tests for selected MOS (ARI Technical Report 757). (AD A192 211)

DuBois, P. H. (1954). A test-dominated society: China, 1115 B.C.-1950 A.D. In _Proceedings, ETS Invitational Conference on Testing_.

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (eds.). (1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1984 fiscal year (ARI Technical Report 660). (AD A178 444)

Flanagan, J. C. (1948). The Aviation Psychology Program in the Army Air Forces, _AAF Aviation Psychology Program Research Reports, 1_, U.S. Government Printing Office.

Flanagan, J. C. (1954). The critical incident technique. _Psychological Bulletin, 51_, 327-358.

Fleishman, E. A. (1967). Performance assessment based on an empirically derived task taxonomy. _Human Factors, 9_, 1017-1032.

Gast, I. F., Campbell, C. H., Steinberg, A. G., & McGarvey, D. A. (1987, August). _A task-based approach for identifying junior NCOs' key responsibilities_. Paper presented at the Annual Convention of the American Psychological Association, New York.

Guilford, J. P. (1957). A revised structure of intelligence (Report No. 19). University of Southern California Psychological Laboratory.

Holland, J. L. (1966). The psychology of vocational choice. Waltham, MA: Blaisdell.

Horst, P. (1954). A technique for the development of a differential prediction battery. _Psychological Monographs_, No. 380.

Horst, P. (1955). A technique for the development of a multiple absolute prediction battery. _Psychological Monographs_, No. 390.

Hough, L. M. (1984). Identification and development of temperament and interest constructs and inventories for predicting job performance of Army enlisted personnel. Minneapolis, MN: Personnel Decisions Research Institute.

Hough, L. M. (ed.). (1986). Literature review: Utility of temperament, biodata, and interest assessment for predicting job performance (ARI Research Note 88-02). (AD A192 109)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report (ARI Research Report 1347). (AD A141 807)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Project A Research Plan (ARI Research Report 1332). (AD A129 728)

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal analysis: Assumptions, models, and data. Beverly Hills, CA: Sage.

Jensen, A. R. (1982). Reaction time and psychometric g. In M. J. Eyesenck (ed.), A model for intelligence. Springer-Verlag.

Joreskog, K. C., & Sorbom, D. (1981). LISREL VI: Analysis of linear squares methods. Uppsala, Sweden: University of Uppsala.

Lord, P., & Novick, M. (1968). Statistical theory of mental test scores. Reading, MA: Addison-Wesley.

McHenry, J. J., & Rose, S. R. (1986). Literature review: Validity and potential usefulness of psychomotor ability tests for personnel selection and classification (ARI Research Note 88-13). (AD A193 558)

Nord, R. D., & White, L. A. (1988). The measurement and application of performance utility. In B. Green, H. Wing, & A. Wigdor (eds.), Linking Military Enlistment Standards to Job Performance, pp. 215-243. Washington, DC: National Academy Press.

Nord, R. D., & White, L. A. (1990). Performance utility and optimal job assignment (ARI Technical Report 904). (AD A229 106)

Peterson, N. (ed.). (1986). Development and field test of the Trial Battery for Project A (ARI Technical Report 739). (AD A184 575)

Peterson, N. G., & Bownas, D. A. (1982). Skills, task structure, and performance acquisition. In M. D. Dunnette and E. A. Fleishman (eds.), Human performance and productivity (Vol. I). Hillsdale, NJ: Lawrence Erlbaum Associates.

Peterson, N., Hough, L., Ashworth, S., & Toquam, J. (1986, November). New predictors of soldier performance. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT.

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. A., Houston, J. S., Toquam, J. L., & Wing, H. (1987). Identification of predictor constructs and development of new selection/classification tests. Paper presented at the Conference of the Society for Industrial and Organizational Psychology, Atlanta.

Peterson, N. G., & Houston, J. S. (1980). The prediction of correctional officer job performance: Construct validation in an employment setting. Minneapolis, MN: Personnel Decisions Research Institute.

Peterson, N. G., Houston, J. S., Rosshardt, M. J., & Dunnette, M. D. (1977). A study of the correctional officer job at Marion Correctional Institution, Ohio: Development of selection procedures, training recommendations and an exit information program. Minneapolis, MN: Personnel Decisions Research Institute.

Peterson, N. G., Houston, J. S., & Rosse, R. L. (1984). The LOMA job effectiveness prediction system: Validity analyses (Technical Report No. 4). Atlanta: Life Office Management Association.

Pulakos, E. D., & Borman, W. C. (eds.). (1986). Development and field test of Army-wide rating scales and the rater orientation and training program (ARI Technical Report 716). (AD B112 857)

Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (1985). The development of administrative measures as indicators of soldier effectiveness (ARI Technical Report 754). (AD A191 232)

Ross, R. T. (1934). Optimum orders of presentation of pairs in pair comparisons. Journals of Educational Psychology, 25, 375-382.

Sadacca, R., Campbell, J. P., White, L. A., & DiFazio, A. S. (1988). Weighting criterion components to develop composite measures of job performance (ARI Technical Report 838). (AD A210 357)

Sadacca, R., White, L. A., Campbell, J. P., DiFazio, A. S., & Schultz, S. R. (1988). Assessing the utility of MOS performance levels in Army enlisted occupations (ARI Technical Report 839). (AD A211 602)

Schmidt, F. L., Hunter, J. E., Croll, P. R., & McKenzie, R. C. (1983). Estimation of employment test validities by expert judgment. _Journal of Applied Psychology, 68_, 590-601.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criterion: A review and resolution of the controversy. _Personnel Psychology, 24_, 419-434.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. _Journal of Applied Psychology, 47_, 149-155.

Staff, AGO, Personnel Research Branch. (1943a). Personnel research in the Army, I. Background and organization, _Psychological Bulletin, 40_, 129-135.

Staff, AGO, Personnel Research Branch. (1943b). Personnel research in the Army, II. The classification system and the place of testing. _Psychological Bulletin, 40_, 205-211.

Staff, AGO, Personnel Research Branch. (1943c). Personnel research in the Army, III. Some factors affecting research in the Army. _Psychological Bulletin, 40_, 237-278.

Staff, AGO, Personnel Research Branch. (1943d). Personnel research in the Army, IV. The selection of radiotelegraph operators. _Psychological Bulletin, 40_, 357-371.

Staff, AGO, Personnel Research Branch. (1943e). Personnel research in the Army, V. The Army specialized training program. _Psychological Bulletin, 40_, 429-435.

Staff, AGO, Personnel Research Branch. (1943f). Personnel research in the Army, VI. The selection of truck drivers. _Psychological Bulletin, 40_, 499-508.

Sternberg, S. (1966). High speed scanning in human memory. _Science, 153_, 652-654.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. _Acta Psychologica, 30_, 276-315.

Stuit, D. B. (1947). Personnel research and test development in the Bureau of Naval Personnel. Princeton, NJ: Princeton University Press.

Thorndike, R. L. (1949). _Personnel selection_. New York: Wiley.

Toquam, J. L., Corpe, V. A., & Dunnette, M. D. (1986). Literature review: Cognitive abilities--theory, history, and validity (ARI Research Note in preparation).

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (1986). Development and field test of behaviorally anchored rating scales for nine MOS (ARI Technical Report 776). (AD A194 271)

Torgerson, W. S. (1958). Theory and methods of scaling. New York: John Wiley & Sons.

White, L. A., Gast, I. G., & Rumsey, M. G. (1986). Categories of leaders' behavior that influence the performance of enlisted soldiers, ARI Working Paper RS-WP-86-1.

Young, W. Y., Harris, J. H., Hoffman, R. G., Houston, J. S., & Wise, L. L. (1987, April). Large scale data collection and data base preparation. Paper presented at the Conference of the Society of Industrial and Organizational Psychology, Atlanta. Personnel Psychology, 43, 2, Summer 1990.

Appendix A
CHARACTERISTICS OF ARMY PERSONNEL SYSTEM
(Described as of February 1989)

The major stages of the selection, classification, and assignment process for persons entering enlisted service in the Army are presented in Table A-1. The size, diversity, and widespread geographical distribution of Army activities have long dictated that the initial stages of personnel recruitment, selection, classification, and training be performed across many specialized units or activities and by personnel who have been specifically trained for these functions with guidance from command. Certain other functions are both formalized and carried out at the command level. These include unit or on-the-job training; performance evaluation; and decisions (or recommendations) concerning promotion, discipline, reassignment, and retention or separation from service. The major stages of the process as of February 1989 are discussed below.

Recruitment

It is difficult to discuss recruitment, selection, and classification separately. They are interdependent processes. Their complementary nature should be evident in the ensuing discussion.

The Army has succeeded in meeting or approximating its numerical recruitment quotas in most of the years following the change to an All-Volunteer Force, resulting in an annual average of about 120,000-140,000 enlisted accessions from over twice as many applicants in the preceding 10 fiscal years. Furthermore, many qualified applicants do not enter active duty immediately but enter the delayed entry program (DEP) where they await a training slot.

The Army seeks to recruit the most capable personnel. Quality is generally defined in terms of high school graduation status and average or above scores on the Armed Forces Qualification Test (AFQT). The AFQT is a composite of four subtests (comprising verbal and math content) from the overall selection and classification instrument, the Armed Services Vocational Aptitude Battery (ASVAB). AFQT scores are reported in percentiles relative to the national youth population. For convenience, they are grouped into the following categories and subcategories:

AFQT Category	Percentile Score Range
I	93 - 100
II	65 - 92
IIIA	50 - 64
IIIB	31 - 49
IVA	21 - 30
IVB	16 - 20
IVC	10 - 15
V	1 - 9

Table A-1

The Army Selection, Classification, and Evaluation Process

Stage/Activity	Process	Outcome
Recruitment (U.S. Army Recruiting Command)	o Recruiting Incentives, Options o Recruiter Interviews o Aptitude Pre-Screening Test (EST) (CAST) o Records Checks	o To MET Sites or MEPS o Disqualified
Selection/ Classification (MEPS)	o Aptitude Testing (ASVAB) o Physical Exam (PULHES) o Moral Screening o Special Tests o Skill/Training Counseling o Classification	o To Training Center o Disqualified
Entry Training (Army Training Centers & Schools)	o Basic Combat Training o Individual Training o Training Evaluation o Assignment o Disciplinary Reviews o Special Courses (NRI, etc.)	o To units o Reassigned/ Recycled o Discharged
First Term (Operating Units)	o Unit (on-the-job) Training and Mission Activities o Special Courses (NRI, etc.) o Evaluation-SQT Ratings, Disciplinary Reviews o Promotion Eligibility o Reenlistment Counseling and Screening o Army Continuing Education System	o Promotion/ Demotion o Discharged (prior to ETS) o Separation (ETS) o Reenlistment
Second Term (Operating Units)	o Unit Training and Mission Activity o Advanced Technical/ Leadership Training o Evaluation o Promotion Eligibility	o Promotion/ Demotion o Reassigned o Discharged (prior to ETS) o Separation (ETS) o Reenlistment

Categories I and II signify well-above and above average trainability, respectively. Category III denotes average trainability, and Category IV signifies below average trainability. Individuals scoring within Category V are, by law, ineligible for enlistment. Because of their likelihood of success in training (and now with evidence of the AFQT's relationship to job performance), the Army attempts to maximize the recruitment of those scoring within Categories I through IIIA. In addition, because traditional high school graduates are more likely to complete their contracted enlistment terms, in contrast to nongraduates and alternative credential holders (e.g., GED credential holders), they are most actively recruited.

Though qualification for initial enlistment into the Army is based upon a number of criteria (including age, moral standards, and physical standards), education and particularly aptitude are the criteria that are most pervasive and most scrutinized. The Army tries to target its advertising and aim its recruiting resources so as to attract quality recruits. As a means of identifying recruitment prospects, while offering a career guidance tool, the ASVAB is administered to 900,000 high school juniors and seniors annually as part of the DoD Student Testing Program.

In order to meet numerical requirements and budget constraints, the Army has recruited some non-high school graduates and applicants scoring in AFQT Category IV. And, between 1976 and 1980, as a result of the ASVAB misnorming the Army erroneously enlisted high proportions of these less-preferred recruits. This situation raised concerns in Congress, and led to the imposition of ceilings on the proportion of non-high school graduates and Category IVs who may be enlisted. One of the outcomes of Project A will be a much more solid empirical basis for qualification decisions. In fact, this research is particularly timely, given indications that banner recruiting times have tapered off.

To compete with the other Services and with the private sector for the prime target group, the Army has had to offer a variety of special inducements, including "critical skill" bonuses and educational incentives. A popular inducement has been the "training of choice" enlistment to a specific school training program, provided that applicants meet the minimum aptitude and educational standards and other prerequisites, and that training "slots" are available at the time of their scheduled entry into the program. Additional options, offered separately or in combination with "training of choice," include guaranteed initial assignment to particular commands, units, or bases, primarily in the combat arms or in units requiring highly technical skills. In recent years, a large proportion of all Army recruits, particularly in the preferred aptitude and educational categories, has been enlisted under one or more of these options. An important research contribution would be to provide counselors with improved data-based aids to help create optimal person-job choices in light of Army manpower needs.

The importance of aptitude in recruiting decisions is exemplified in the prescreening of applicants at the recruiter level. For applicants who have not previously taken the ASVAB and whose educational/aptitude qualifications appear to be marginal based on the Army's trainability standards, the recruiter may administer a short Computerized Adaptive Screening Test (CAST) or Enlisted Screening Test (EST) to assess the applicant's prospects of passing the ASVAB. The Army has also employed non-cognitive tests to identify individuals who are likely to be poor risks in terms of the probability of

A-3

completing of Army basic training. Applicants who appear, upon initial recruiter screening, to have a reasonable chance of qualifying for service are referred either to one of 759 Mobile Examining Team (MET) sites for administration of the ASVAB, or directly to a Military Entrance Processing Station (MEPS) where all aspects of enlistment testing are conducted.

Selection and Classification at the MEPS

Based on the information assembled, classification and assignment to a particular training activity are completed at the MEPS for applicants found qualified for enlistment.

The current versions of the ASVAB (Forms 11-13) consist of the following 10 subtests:

1. Arithmetic Reasoning
2. Numerical Operations
3. Paragraph Comprehension
4. Word Knowledge
5. Coding Speed
6. General Science
7. Mathematics Knowledge
8. Electronics Information
9. Mechanical Comprehension
10. Automotive-Shop Information

In addition to AFQT scores, subtest scores are combined to form 10 aptitude composite scores, based on those combinations of subtests that have been found to be most valid as predictors of successful completion of the various Army school training programs. For example, the composite score for administrative specialties is based on the numerical operations, paragraph comprehension, word knowledge, and coding speed subtests. The composite score for electronics specialties is based on a combination of the scores for arithmetic reasoning, general science, mathematics knowledge, and electronics information.

As stated above, eligibility for enlistment, in terms of the trainability standard, is based upon a combination of criteria: AFQT score, aptitude area composite scores, and whether the applicant is or is not a high school diploma graduate. Under the most recent Army regulation[1], the following standards were in effect:

- High school graduates are eligible if they achieve an AFQT percentile score of 16 or higher and a standard score of 85 in at least one aptitude area.

- GED high school equivalency holders are eligible if they achieve an AFQT percentile score of 31 or higher and a standard score of 85 in at least one aptitude area.

---

[1]Army Regulation 601-201, 1 October 1980, revised, Table 2-2.

- Non-high school graduates are eligible only if they achieve an AFQT percentile score of 31 or higher and standard scores of 85 in at least two aptitude areas.

Physical standards are captured in the PULHES profile, which rates the applicant on General Physical (P), Upper torso (U), Lower torso (L), Hearing (H), Eyes (E), and Psychiatric. The Army also sets general height and weight standards for enlistment.

Initial Classification

The overwhelming majority of Army enlistees enter the Army under a specific enlistment option that guarantees choice of initial school training, career field assignment, unit assignment, or geographical area. For these applicants, the initial classification and training assignment decision must be made prior to entry into service. This is accomplished at MEPS by referring applicants who have passed the basic screening criteria (aptitude, physical, moral) to an Army guidance counselor, whose responsi-bility is to match the applicant's qualifications and preferences to Army current skill training requirements, and to make "reservations" for training assignments, consistent with the applicant's enlistment option.

For the enlistee, this decision will determine the nature of his or her initial training and occupational assignment, future military work environ-ment, and chances of successful advancement in an Army career. For the Army, the relative success of the assignment process will significantly determine the aggregate level of performance and attrition for the entire force.

The classification and training "reservation" procedure is accomplished by the Recruit Quota System (REQUEST) which was implemented in 1973. REQUEST is a computer-based system designed to coordinate the information needed to reserve training slots for volunteers. REQUEST uses minimum qualifications for accessions control. Thus, to the extent that an applicant may minimally qualify for a wide range of courses or specialties, based on aptitude test scores, the initial classification decision is governed by (a) his or her own stated preference (often based upon limited knowledge about the actual job content and working conditions of the various military occupations), (b) the availability of training slots, and (c) the current priority assigned to filling each military occupational specialty (MOS).

These interactions among recruitment, selection, and classification in the current Army system give rise to several issues. First, there is an evident need for decision-making algorithms designed to maximize the overall utility of the MOS assignments. This requires that the average differential utilities of alternative assignments be known, as well as the marginal utility of each additional assignment to an MOS. The Army system currently incorporates marginal utilities by specifying desired distributions of AFQT scores, which are termed quality goals. In general, the parameters of recruit supply and demand (e.g., number of applicants in various categories, selection ratio, percentage of training slots filled, MOS priority) must also be taken into account when developing decision-making algorithms for selection and classification. The decision process must also allow for the potentially adverse impacts on recruitment if the enlistee's interests, work values, and preferences are not given sufficient weight. There are clear trade-offs that

must be evaluated between the procedures necessary to (a) attract qualified people, and (b) put them into the right slots.

Initial Training

After processing at a Reception Battalion, all non-prior service Army recruits are assigned to a basic training program (BCT) of 8 weeks which is followed, with few exceptions, by a period of advanced individual training (AIT), designed to provide basic entry-level skills. Entrants into the combat arms and the military police receive both their basic training and their AIT at the same Army base (One Station Unit Training) in courses of about 3-4 months' total duration. Those assigned to other specialties are sent to separate Army technical schools whose course lengths vary considerably, depending upon the technical complexity of the MOS. The diversity of course offerings is illustrated by the fact that the Army provides initial skills training in about 240 separate courses.[*]

In contrast to earlier practice, most enlisted trainees do not currently receive school grades upon completion of their courses, but are evaluated under Pass/Fail criteria. Those initially failing certain portions of a course are recycled. The premise is that slower learners, given sufficient time and effort under self-paced programs, can normally be trained to a satisfactory level of competence, and that this additional training investment is cost-effective. Those who continue to fail the course may be reassigned to other, often less demanding specialties or discharged from service. One consequence of these practices is to limit the usefulness of the selection/ classification practices as predictors of later performance.

Performance Assessment in Army Units

Upon assignment to an Army unit, most of the personnel actions affecting the career of the first-term enlistee are initiated by his or her immediate supervisor and/or the unit commander. These include the nature of the duty assignment, the provision of on-the-job or unit training, and assessments of performance, both on and off the job. These assessments in-fluence such decisions as promotion, future assignment, and eligibility for reenlistment, as well as possible disciplinary action (including early discharges from service).

To assure that these processes are administered fairly and consistently, in a manner compatible with broader Army objectives, the various aspects of enlisted personnel management are governed by detailed Army regulations. Army Regulation 600-211, The Enlisted Personnel Management System, and related regulations cover such subjects as enlisted personnel evaluation and promotion, while AR 601-280, The Army Reenlistment Program, prescribes the qualifications for reenlistment.

During an initial 3-year enlistment term, the typical enlistee can expect to progress to pay grade E-4, although advancement to higher pay grades for specially qualified personnel is not precluded. Authority to promote qualified personnel up to grade E-4 is delegated to unit commanders; promotion

---

[*]Department of Defense, Military Manpower Training Report for 1982, March 1981, p. II-4.

to higher grades is numerically restricted and must be approved either by field grade commanders for grades E-5 and E-6 or by HQDA for grades E-7 through E-9. Promotion to E-2 is almost automatic after 6 months of service. Promotions to grades E-3 and E-4 normally require completion of certain minimum periods of service (12 and 24 months, respectively), but are subject to certain numerical strength limitations and specific commander approval. Unit commanders also have the authority to reduce assigned soldiers in pay grade, based on misconduct or inefficiency.

The Enlisted Evaluation System provides for an evaluation both of the soldier's proficiency in his or her MOS and of overall duty performance. The process includes a subjective evaluation based on supervisory performance appraisal and ratings that are conducted at the unit level under prescribed procedures, and an objective evaluation based on the results of a Skill Qualification Test (SQT). The latter is a criterion-referenced, paper-and-pencil performance-knowledge test which evaluates the soldier's ability to perform critical job tasks satisfactorily. The responsibility for planning and developing the SQT and of validating its results lies with the U.S. Army Training Support Center of the Training and Doctrine Command (TRADOC); actual administration of the tests has been delegated to each of the major Army commands.

The current SQTs are developed primarily by individuals (e.g., enlisted personnel, officers, and civilians) who are knowledgeable about task elements and performance requirements but are not trained as test designers.

Reenlistment Screening

The final stage of personnel processing of first-term enlisted personnel is screening for reenlistment eligibility which, as described in AR 601-280, considers such criteria as disciplinary records; aptitude area scores (based on ASVAB or its predecessors); low SQT scores, when applicable; and slow grade progression "resulting from a pattern of marginal conduct and/or performance." Enlisted personnel who do not meet certain minimum standards under these criteria must be approved by Commanding General of the Personnel Command, before being processed for reenlistment.

The cumulative losses due to attrition, reenlistment screening, and non-reenlistment of eligible personnel have resulted in the progressive diminution of initial Army cohorts to about 20-30 percent of their original numbers by the time they enter the fourth year of enlisted service. Moreover, not all of the group that remains are retained or wish to be retained in their original specialties, since an offer of retraining is often an inducement for reenlistment. The cumulative impact of this skill drain upon the Army is considerable.

Summary

Even this brief description of the system illustrates the complexity of the Army's personnel decision-making requirements and the large number of parameters that must be taken into account. In addition, decisions must be made for a very large flow of individuals within a very short time frame. In this regard the Army faces a much more difficult personnel management task than virtually any other organization. More effective selection/classification/promotion strategies would pay large dividends.

Appendix B

BIBLIOGRAPHY OF PROJECT A PUBLICATIONS

Fiscal Year 1983

Borman, W. C., Motowidlo, S. J., & Hanser, L. M. (1983, August). Developing a model of soldier effectiveness: A strategy and preliminary results. Paper presented at the Annual Convention of the American Psychological Association, Anaheim, CA. (In ARI Research Note 83-37, October 1983.)

Eaton, N. K., Weltin, M., & Wing, H. (1982). Validity of the Military Applicant Profile for predicting early attrition in different educational, age, and racial groups. (In ARI Research Note 83-37, October 1983; published as ARI Technical Report 567, December 1982.)

Eaton, N. K., Wing, H., & Mitchell, K. (1983, August). Putting the "dollars" into utility analysis. Paper presented at the Annual Convention of the American Psychological Association, Anaheim, CA. (In ARI Research Note 83-37, October 1983.)

Friedman, D., Streicher, A. H., Wing, H., & Grafton, F. (1982, November). Assessment of practice effects: Test-retest scores for FY81 active Army applicants on ASVAB 8/9/10. Paper presented at the Annual Conference of the Military Testing Association, San Antonio. (In ARI Research Note 83-37, October 1983.)

Grafton, F. C., Mitchell, K. J., & Wing, H. (1983). Final status report on the comparability of ASVAB 6/7 and 8/9/10 Aptitude Area score scales. (In ARI Research Note 83-37, October 1983; issued as ARI Selection and Classification Technical Area Working Paper 82-7, March 1983).

Hanser, L. M., & Grafton, F. C. (1983, August). Dusting off old data: Encounters with archival records. Paper presented at the Annual Convention of the American Psychological Association, Anaheim, CA. (In ARI Research Note 83-37, October 1983.)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1983, May). Improving the selection, classification, and utilization of Army enlisted personnel: Project A Research Plan. (ARI Research Report 1332.)

Mitchell, K. J. (1983). Verbal information processing paradigms: A review of theory and methods. (In ARI Research Note 83-37, October 1983; published as ARI Technical Report 648, September 1984.)

Osborn, W. (1983, August). Issues and strategies in measuring performance in Army jobs. Paper presented at the Annual Convention of the American Psychological Association, Anaheim, CA. (In ARI Research Note 83-37, October 1983.)

Oxford-Carpenter, R., & Schultz, L. J. (1983). Reading assessment in the Army. (In ARI Research Note 83-37, October 1983; issued as ARI Selection and Classification Technical Area Working Paper 83-5, November 1983.)

Oxford-Carpenter, R. L., & Schultz, L. J. (1983). Preliminary thoughts about research on reliability and validity of Army training measures. (In ARI Research Note 83-37, October 1983; issued as ARI Selection and Classification Technical Area Working Paper 83-7, November 1983.)

Weltin, M. M., & Popelka, B. A. (1983). Evaluation of the ASVAB 8/9/10 Clerical Composite for predicting training performance. (In ARI Research Note 83-37, October 1983; published as ARI Technical Report 594, October 1983.)

Wetrogan, L. I., Olson, D. M., & Sperling, H. M. (1983, August). Job performance and assessment: A systemic model. Paper presented at the Annual Convention of the American Psychological Association, Anaheim, CA. (In ARI Research Note 83-37, October 1983.)

Wise, L. L., Wang, M., & Rossmeissl, P. Longitudinal research database plan. (1983). (In ARI Research Note 83-37, October 1983; published as ARI Research Report 1346, December 1983.)

Fiscal Year 1984

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1984).
Development of a model of soldier effectiveness. (In ARI Technical
Report 660, July 1985; appendixes in ARI Research Note 85-14, October
1984.)

Borman, W. C., White, L. A., & Gast, I. F. (1984, August). Factors
relating to peer and supervisor ratings of job performance. Paper
presented at the Annual Meeting of the American Psychological
Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

Brandt, D., McLaughlin, D., Wise, L., & Rossmeissl, P. (1984, August).
Complex cross-validation of the validity of a predictor battery. Paper
presented at the Annual Convention of the American Psychological
Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

Brandt, D., McLaughlin, D. H., Wise, L. L., & Rossmeissl, P. G. (1984,
August). Adjustments for the effects of range restriction on composite
validity. Paper presented at the Annual Convention of the American
Psychological Association, Toronto, Canada. (In ARI Technical Report
660, July 1985.)

Eaton, N. K. (1984, May). The U. S. Army research project to improve
selection and classification decisions. Paper presented at the National
Security Industrial Association Conference on Personnel and Training
Factors in Systems Effectiveness, Springfield, VA.

Eaton, N. K., & Goer, M. H. (Eds.). (1983, October). Improving the
selection, classification, and utilization of Army enlisted personnel:
Technical Appendix to the Annual Report. (ARI Research Note 83-37.)

Eaton, N. K., Wing, H., & Mitchell, K. J. (1984). Alternate methods of
estimating the dollar value of performance. Personnel Psychology, 38,
27-40, 1985. (In ARI Technical Report 660, July 1985.)

Hanser, L. M., & Mitchell, K. J. (1983, October). Factorial invariance of
the Armed Services Vocational Aptitude Battery. Paper presented at the
Annual Conference of the Military Testing Association, Gulf Shores, AL.
(In ARI Research Note 83-37, October 1983.)

Hough, L. M. (1984). Identification and development of temperament and
interest constructs and inventories for predicting job performance of
Army enlisted personnel. Minneapolis, MN: Personnel Decisions Research
Institute.

Hough, L., Dunnette, M. D., Wing, H., Houston, J., & Peterson, G. (1984,
August). Covariance analyses of cognitive and noncognitive measures of
Army recruits: An initial sample of preliminary battery. Paper
presented at the Annual Convention of the American Psychological
Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1983, October). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report. (ARI Research Report 1347.)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1984, August). Selecting job tasks for criterion tests of MOS Proficiency.

Martin, C. J., Rossmeissl, P. G., & Wing, H. (1983, November). Validity of cognitive tests in predicting Army training success. Paper presented at the Psychonomics Society, San Diego. (In ARI Technical Report 660, July 1985.)

McLaughlin, D. (1984, August). Differential validity of ASVAB for job classification. Paper presented at the Annual Convention of the American Psychological Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D.A., & Wang, M. (1984, May). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY1981 and FY1982 enlisted accessions. (ARI Technical Report 651.)

Olson, D. M., Borman, W. C., Roberson, L., & Rose, S. R. (1984, August). Relationships between scales on an Army work environment questionnaire and measures of performance. Paper presented at the Annual Convention of the American Psychological Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

Olson, D. M., & Hanser, L. M. (1983, October). Examination of ability requirements for the Infantry Career Management Field. Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, AL. (In ARI Research Note 83-37, October 1983.)

Osborn, W., & Hoffman, R. G. (1984, August). The cost-effectiveness of hands-on and knowledge measures. Paper presented at the Annual Convention of the American Psychological Association, Toronto, Ontario. (In ARI Technical Report 660, July 1985.)

Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (1984, August). Administrative records as effectiveness criteria: An alternative approach. Paper presented at the Annual Convention of the American Psychological Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

Rosse, R. L., Borman, W. C., Campbell, C. H., & Osborn, W. C. (1983, October). Grouping Army occupational specialties by judged similarity. Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, AL. (In ARI Research Note 83-37, October 1983.)

Rossmeissl, P. G., & Brandt, D. A. (1984 August). Subgroup variation in the validity of Army aptitude area composites. Paper presented at the Annual Convention of the American Psychological Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

Rossmeissel, P. G., & Eaton, N. K. (1984, April). An analysis of SQT scores as a function of aptitude area composite scores for logistics MOS. (In ARI Technical Report 660, July 1985.)

Rossmeissl, P. G., Martin, C. J., Wing, H., & Wang, M. (1983, October). Validity of ASVAB 8/9/10 for predicting training success. Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, AL. (In ARI Research Note 83-37, October 1983.)

Rossmeissl, P. G., & Stern, B. M. (1983, November). The application of meta-analytic techniques in estimating selection/classification parameters. Paper presented at the Psychonomics Society, San Diego. (In ARI Technical Report 660, July 1985.)

Rossmeissl, P. G., Wise, L. L., & Wang, M. (1983, October). A data base system for validation research. Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, AL. (In ARI Technical Report 660, July 1985.)

Wing, H. (1984, August). Meta-analysis: Procedures, practices, pitfalls-- Introductory remarks. Presented at the Annual Convention of the American Psychological Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

Wing, H., Peterson, N. G., & Hoffman, R. G. (1984, August). Expert judgments of predictor-criterion validity relationships. Paper presented at the Annual Convention of the American Psychological Association, Toronto, Canada. (In ARI Technical Report 660, July 1985; appendixes in ARI Research Note 85-14, October 1984.)

Wise, L. L., McLaughlin, D. H., Rossmeissl, P. G., & Brandt, D. A. (1984, August). Clustering military occupations in defining selection and classification composites. Paper presented at the Annual Convention of the American Psychological Association, Toronto, Canada. (In ARI Technical Report 660, July 1985.)

<u>Fiscal Year 1985</u>

Borman, W. C. (1985). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army officer sample. (In ARI Research Note 87-54.)

Borman, W. C. (1985, August). "Personal constructs and 'folk theories' of subordinate performance: Cognitive psychology contributions to performance rating research." In Borman, W. C. & Kane, J. S. <u>Performance measurement and appraisal: Some theoretical considerations and their implications for practice.</u> Presented as part of American Psychological Association Convention Division 14 Workshop.

Borman, W. C., White, L. A., Gast, I. F., & Pulakos, E. D. (1985 August). <u>Performance rating as criteria: What is being measured?</u> Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Campbell, J. P. & Harris, J. H. (1985, August). <u>Criterion reduction and combination via a participative decision-making panel.</u> Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Campbell, R. C. (1985). <u>Scorer training materials.</u> (ARI Working Paper RS-WP-85).

Eaton, N. K. (1985, August). <u>Measurement of entry-level job performance.</u> Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (Eds.) (1985, July). <u>Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1984 fiscal year.</u> (ARI Technical Report 660.)

Hough, L. M., Barge, B. N., Houston, J. S., McGue, M. K., & Kamp, J. D. (1985, August). <u>Problems, issues, and results in the development of temperament, biographical, and interest measures.</u> Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1985, July). <u>Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report synopsis, 1984 fiscal year.</u> (ARI Research Report 1393.)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1984, October). <u>Improving the selection, classification, and utilization of Army enlisted personnel: Appendixes to the Annual Report, 1984 fiscal year.</u> (ARI Research Note 85-14.)

McHenry, J. J., & McGue, M. K. (1985, August). <u>Problems, issues, and results in the development of computerized psychomotor measures</u>. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

McLaughlin, D. H. (1985, August). <u>Measurement of test battery value for selection and classification</u>. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Peterson, N. G. (1985, August). <u>Overall strategy and methods for expanding the measured predictor space</u>. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (1985). <u>The development of administrative measures as indicators of soldier effectiveness</u>. (Published as ARI Technical Report 754, August 1987.)

Rosse, R. L., & Peterson, N. (1985, August). <u>Advantages and problems with using portable computers for personnel measurement</u>. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Rossmeissl, P. G., & Brandt, D. A. (1985, August). <u>Modeling the selection process to adjust for restriction in range</u>. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985, August). <u>Comparing work sample and job knowledge measures</u>. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Sadacca, R., & Campbell, J. P. (1985, March). <u>Assessing the utility of a personnel/classification system</u>. Paper presented at the Meeting of the Southeastern Psychological Association, Atlanta, GA. (In ARI Research Note 87-54.)

Toquam, J. L., Dunnette, M. D., Corpe, V., McHenry, J. J., Keyes, M. A., McGue, M. K., Houston, J. S., Russell, T. L., & Hansen, M. A. (1985, August). <u>Development of cognitive/perceptual measures: Supplementing the ASVAB</u>. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Walker, C. B. (1984, November). <u>Validation of the Army's Military Applicant Profile (MAP) against an expanded criterion space</u>. Paper presented at the meeting of the Military Testing Association, Munich, Germany. (In ARI Research Note 87-54.)

Walker, C. B. (1985, February). <u>The fakability of the Army's Military Applicant Profile (MAP)</u>. Paper presented at the Combined National and Western Region Meeting of the Association of Human Resources Management and Organizational Behavior, Denver. (In ARI Research Note 87-54.)

White, L. A., Gast, I. F., Sperling, H. M., & Rumsey, M. G. (1984, November). _Influence of soldiers' experiences with supervisors on performance during the first tour._ Paper presented at the meeting of the Military Testing Association, Munich, Germany. (In ARI Research Note 87-54.)

Wing, H. (1985, August). _Expanding the measurement of predictor space for military enlisted jobs._ Symposium presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Wise, L. L., & Mitchell, K. J. (1985, August). _Development of an index of maximum validity increment for a new predictor measures._ Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note 87-54.)

Arabian, J., Rumsey, M., & McHenry, J. (1986, September). Army research to link standards for enlistment to on-the-job performance. (ARI Selection and Classification Technical Area Working Paper.)

Arabian, J. M., & Hanser, L. M. (1986, August). Standard setting procedures: Army enlistment standards and job performance. Paper presented at the Annual Convention of the American Psychological Association, Washington, DC. (In ARI Research Note 813704.)

Borman, W. C. (1986, August). Performance criterion measurement: What are the different methods measuring? Paper presented at the Air Force Conference on Job Performance Measurement, Air Force Human Resources Laboratory, San Antonio. (In ARI Research Note 813704.)

Borman, W. C. (1986, July). Measuring soldier performance in the U.S. Army. Invited paper presented to the Canadian Forces Personnel Applied Research Unit, Toronto.

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1986). Development of a model of soldier effectiveness. (Published as ARI Technical Report 741, May 1987.)

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1986). Development of a model of soldier effectiveness: Retranslation materials and results. (ARI Research Note 87-29, May 1987.)

Borman, W. C., Pulakos, E. D., & Motowidlo, S. J. (1986, August). Toward a general model of soldier effectiveness. Paper presented at the Annual Convention of the American Psychological Association, Washington, DC. (In ARI Research Note 813704.)

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986, July). Development and field test of task-based MOS-specific criterion measures. (ARI Technical Report 717.)

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986). Appendixes to ARI Technical Report 717: Development and field test of task-based MOS-specific criterion measures. Issued in volumes as follows:

> Volume 1, Appendixes A-E, ARI Research Note 88-15, April 1988.
> Volume 2, Appendixes F and G (Part 1), ARI Research Note 88-16, April 1988.
> Volume 3, Appendix G (Part 2), ARI Research Note 88-19, April 1988.
> Volume 4, Appendix H (Part 1), ARI Research Note 88-20, April 1988.
> Volume 5, Appendixes H (Part 2), I-J, ARI Research Note 88-21, April 1988.
> Volume 6, Appendixes K-O, ARI Research Note 88-22, April 1988.
> Volume 7, Appendixes P-U, V (Part 1), ARI Research Note 88-23, April 1988.

Volume 8, Appendix V (Part 2), ARI Research Note 88-24,
April 1988.
Volume 9, Appendix V (Part 3), ARI Research Note 88-25,
April 1988.
Volume 10, Appendix V (Part 4), ARI Research Note 88-26,
April 1988.

Campbell, J. P. (1986, August). Project A: When the textbook goes
operational. Paper presented at the Annual Convention of the American
Psychological Association, Washington, DC. (In ARI Research Note
813704.)

Campbell, J. P. (Ed.). (1986). Improving the selection, classification,
and utilization of Army enlisted personnel: Annual Report, 1985 fiscal
year. (Published as ARI Technical Report 745, June 1987.)

Davis, R. H., Davis, G. A., Joyner, J. N., & de Vera, M. V. (1986).
Development and field test of job-relevant knowledge tests for selected
MOS. (Published as ARI Technical Report 757, August 1987.)

Davis, R. H., Davis, G. A., Joyner, J. N., & de Vera, M. V. (1986).
Appendixes to ARI Technical Report 757: Development and field test of
job-relevant knowledge tests for selected MOS. (ARI Research Note in
preparation.)

Eaton, N. K., Wing, H., & Lau, A. (1985, October). Utility estimation in
five enlisted occupations. Paper presented at the Annual Conference of
the Military Testing Association, San Diego. (In ARI Research Note
813704.)

Ford, P., Campbell, C. H., Felker, D. R., & Edwards, D. C. (1986, August).
Comparability of hands-on and knowledge tests across nine military jobs.
Paper presented at the Annual Convention of the American Psychological
Association, Washington, DC. (In ARI Research Note 813704.)

Hanser, L. M., & Arabian, J. M. (1985, October). Multi-dimensional
performance measurement. Paper presented at the Annual Conference of
the Military Testing Association, San Diego. (In ARI Research Note
813704.)

Hough, L. M. (Ed.). (1986). Literature review: Utility of temperament,
biodata, and interest assessment for predicting job performance. (PDRI
Technical Report 116.) (Published as ARI Research Note 88-02, January
1988.)

Hough, L. M., Gast, I. F., White, L. A., & McCloy, R. (1986, August). The
relation of leadership and individual differences to job performance.
Paper presented at the Annual Convention of the American Psychological
Association, Washington, DC. (In ARI Research Note 813704.)

Hough, L. M., McGue, M. K., Kamp, J. D., Houston, J. S., & Barge, B. N. (1985,
October). Measuring personal attributes: Temperament, biodata, and
interests. Paper presented at the Annual Conference of the Military
Testing Association, San Diego. (In ARI Research Note 813704.)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute. (1986). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1985 fiscal year - Supplement to ARI Technical Report 746. (ARI Research Note 87-54.)

Humphreys, L. G. (1986, August). Stability and instability of individual differences. Paper presented at the Annual Convention of the American Psychological Association, Washington, DC. (In ARI Research Note 813704.)

McHenry, J. J., & Rose, S. R. (1986). Literature review: Validity and potential usefulness of psychomotor ability tests for personnel selection and classification. (Published as ARI Research Note 88-13.)

McHenry, J. J., & Toquam, J. L. (1985, October). Computerized assessment of perceptual and psychomotor abilities. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Olson, D. M., & Borman, W. C. (1985, October). Examination of environmental determinants of Army performance criteria. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Olson, D. M., Borman, W. C., & Motowidlo, S. J. (1986, August). Individual differences and environmental determinants of Army performance criteria. Paper presented at the Annual Convention of the American Psychological Association, Washington, DC. (In ARI Research Note 813704.)

Peterson, N. (Ed.). (1986). Development and field test of the Trial Battery for Project A. (Published as ARI Technical Report 739.)

Peterson, N. (Ed.). (1986). Appendixes to ARI Technical Report 739: Development and field test of the Trial Battery for Project A. (ARI Research Note 87-24.)

Peterson, N. G. (1985, October). Mapping predictors to criterion space: Overview. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Pulakos, E. D., & Borman, W. C. (Eds.). (1986). Development and field test of Army-wide rating scales and the rater orientation and training program. (ARI Technical Report 716.)

Pulakos, E. D., & Borman, W. C. (Eds.). (1986). Appendixes to ARI Technical Report 716: Development and field test of Army-wide rating scales and the rater orientation and training program. (Published as ARI Research Note 87-22.)

Rosse, R. L., & Peterson, N. (1985, October). Using microcomputers for assessment: Practical problems and solutions. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Rossmeissl, P. G., McLaughlin, D. H., Wise, L. L., & Brandt, D. A. (1985, October). The validity of ASVAB for predicting training and SQT performance. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Sadacca, R., de Vera, M. V., DiFazio, A., & White, L. A. (1986, August). Weighting performance constructs in composite measures of job performance. Paper presented at the Annual Convention of the American Psychological Association, Washington, DC. (In ARI Research Note 813704.)

Smith, E. P. (1985, October). Developing new attribute requirements scales for military jobs. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Toquam, J. L., Corpe, V. A., and Dunnette, M. D. (1986). Literature review: Cognitive abilities -- theory, history, and validity. (ARI Research Note in preparation.)

Toquam, J. L., Dunnette, M. D., Corpe, V. A., & Houston, J. Adding to the ASVAB: Cognitive paper-and-pencil measures. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (1986). Development and field test of behaviorally anchored rating scales for nine MOS. (ARI Technical Report in preparation.)

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (1986). Appendixes to ARI Technical Report: Development and field test of behaviorally anchored rating scales for nine MOS. (ARI Research Note in preparation.)

Walker, C. B. (1985, October). Three variables that may influence the validity of biodata. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

White, L. A., Borman, W. C., & Hough, L. M. (1986, August). A path analytic model of job performance ratings. Paper presented at the Annual Convention of the American Psychological Association, Washington, DC. (In ARI Research Note 813704.)

White, L. A., Gast, I. F., & Rumsey, M. G. (1985, October). Leaders' behavior and the performance of first-term soldiers. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Wing, H., Barge, B. N., & Hough, L. M. (1985, October). Vocational interests as predictors of Army performance. Paper presented at the Annual Conference of the Military Testing Association, San Diego. (In ARI Research Note 813704.)

Wise, L. L., Campbell, J. P., McHenry, J. J., & Hanser, L. M. (1986, August). A latent structure model of job performance factors. Paper presented at the Annual Convention of the American Psychological Association, Washington, DC. (In ARI Research Note 813704.)

Fiscal Year 1987

Arabian, J. M., & Mason, J. K. (1986, November). Relationship of SQT scores to Project A measures. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Barge, B. N. (1987, August). Characteristics of biodata items and their relationship to validity. Paper presented at the Annual Convention of the American Psychological Association, New York. (In ARI Research Note 88-23.)

Campbell, C. H. (1986). Developing basic criterion scores for hands-on tests, job knowledge tests, and task rating scales (HumRRO IR-PRD-87-15). (ARI Technical Report in preparation.)

Campbell, C. H., Borman, W. C., Felker, D. B., Ford, P., de Vera Park, M. V., Pulakos, E. D., Riegelhaupt, B. J., & Rumsey, M. G. (1987, April). Development of Project A performance measures. Paper presented at the Conference of the Society for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.)

Campbell, C. H., Ford, P., Rumsey, M. G., Pulakos, E. D., Borman, W. C., Felker, D. B., DeVera, M. V., Riegelhaupt, B. J. (1987). Development of multiple job performance measures in a representative sample of jobs. Personnel Psychology, 43, 2, Summer 1990.

Campbell, C. H., & Hoffman, R. G. (1986, October). Hands-on data collection during Concurrent Validation: Lessons learned (HumRRO IR-PRD-87-16). (ARI Technical Report in preparation.)

Campbell, C. H., & Rumsey, M. G. (1986, November). Skill requirement influences on measurement method intercorrelations. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Campbell, J. P. (1986, December). Validation analysis for new predictors. (ARI Selection and Classification Technical Area Working Paper 86-09.)

Campbell, J. P. (1987). An overview of the Army selection and classification project (Project A). Personnel Psychology, 43, 2, Summer 1990.

Campbell, J. P. (Ed.). (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1986 fiscal year (HumRRO IR-PRD-87-10). (ARI Technical Report 813101.)

Campbell, J. P., Hanser, L. M., & Wise, L. (1986, November). The development of a model of the Project A criterion space. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1987, April). <u>Analysis of criterion measures: The modeling of performance</u>. Paper presented at the Conference of the Society for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.) <u>Personnel Psychology, 43</u>, 2, Summer 1990.

Campbell, R. C., Campbell, C. H., & Doyle, E. L. (1986, November). <u>Patterns of Skill Level One performance in representative Army jobs: Common and technical task comparisons</u>. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Ford, P., & Hoffman, R. G. (1986, November). <u>Effects of test programs on task proficiency</u>. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Gast, I. F., Campbell, C. H., Steinberg, A. G., & McGarvey, D. A. (1987, August). <u>A task-based approach for identifying junior NCO's key responsibilities</u>. Paper presented at the Annual Convention of the American Psychological Association, New York. (In ARI Research Note 88-23).

Gast, I. F., & White, L. A. (1986, November). <u>Effects of soldier performance and characteristics on relationships with superiors</u>. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Harris, J. H., Campbell, J. P., & Campbell, C. H. (1986, November). <u>The Project A Concurrent Validation data collection</u>. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Hoffman, R. G. (1986, November). <u>Post differences in hands-on tests</u>. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Hoffman, R. G. (1987, July). <u>Clustering Army Military Occupational Specialties for Project A: Phase Two</u>. (HumRRO IR-PRD-87-22) (ARI Technical Report in preparation.)

Hoffman, R. G., & Ford, P. (1986, November). <u>Estimates of task parameters for test and training development</u>. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Hough, L. M. (1987, August). <u>Overcoming objections to the use of temperament variables in selection</u>. Paper presented at the Annual Convention of the American Psychological Association, New York. (In ARI Research Note 88-23.)

Hough, L. M., & Ashworth, S. D. (1987, April). <u>Predicting soldier performance: Assessment of temperament constructs as predictors of job performance</u>. Paper presented at the Conference for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1986 fiscal year - Supplement to ARI Technical Report 813101 (HumRRO IR-PRD-87-12). (ARI Research Note 813704.)

Kuhn, D. B. (1987). The assignment of knowledge test items to functional and cognitive categories (HumRRO IR-PRD-87-17). (ARI Research Note 88-28.)

McHenry, J. J., Harris, J. H., & Oppler, S. M. (1986, November). Using confirmatory factor analysis to aid in assessing task performance. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1987, April). Project A validity results: The relationship between predictor and criterion domains. Paper presented at the Conference of the Society for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.) Personnel Psychology, 43, 2, Summer 1990.

McHenry, J. J., Wise, L. L., Campbell, J. P., & Hanser, L. M. (1986, December.) A latent structure model of job performance factors: Appendix. (ARI Selection and Classification Technical Area Working Paper 86-10.)

Nord, R., & White, L. A. (1987, August). Optimal job assignment and the utility of performance. Paper presented at the 95th Annual Convention of the American Psychological Association, New York. (In ARI Research Note 88-23.)

Olson, D. M., & Borman, W. C. (1986, November). Influence of environment, ability, and temperament on performance in Army MOS. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Olson, D. M., & Borman, W. C. (1987, May). Development and field tests of the Army Work Environment Questionnaire (ARI Technical Report 737).

Peterson, N., Hough, L., Ashworth, S., & Toquam, J. (1986, November). New predictors of soldier performance. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (in ARI Research Note 88-23.)

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. A., Houston, J. S., Toquam, J. L., & Wing, H. (1987, April). Identification of predictor constructs and development of new selection/classification tests. Paper presented at the Conference of the Society for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.) Personnel Psychology, 43, 2, Summer 1990.

Pulakos, E. D., & Borman, W. C. (1987). Developing the basic criterion scores for Army-wide and MOS-specific ratings. (ARI Technical Report in preparation.)

Pulakos,, E. D., Hanson, M. A., Borman, W. C., Hallam, G., Carter, G., & Owens-Kurtz, C. (1987, August). _Developing behavioral rating scales to evaluate second tour performance_. Paper presented at the 95th Annual Convention of the American Psychological Association, New York. (In ARI Research Note 88-23.)

Pulakos, E. D., White, L. A., & Borman, W. C. (1987, April). _An examination of race and sex effects on performance ratings_. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.)

Radtke, P., & Edwards, D. S. (1986, November). _Effect of practice on soldier task performance_. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Rumsey,, M. G. (1987, August). _Getting answers to the right questions: Job analysis strategy_. Paper presented at the Annual Convention of the American Psychological Association, New York. (In ARI Research Note 88-23.)

Sadacca, R., Campbell, J. P., Wise, L. L., & White, L. A. (1987, April). _Performance composites, performance utility, and selection/classification decisions_. Paper presented at the Conference of the Society for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.)

Sadacca, R., Campbell, J. P., DiFazio, A. S., Schultz, S. R., & White, L. A. (1987). Scaling performance utility to enhance selection/classification decisions. _Personnel Psychology, 43,_ 2, Summer 1990.

Schultz, S. R., Kuhn, D. B., & Walker, C. B. (1987, September). _Development of job-relevant knowledge tests for Longitudinal Validation_ (HumRRO FR-PRD-87-29). (ARI Technical Report in preparation.)

Shields, J. L., & Hanser, L. M. (1987, April). _Designing, planning, and selling Project A_. Paper presented at the Conference of the Society for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.) _Personnel Psychology, 43,_ 2, Summer 1990.

Smith, E. P., & Rossmeissl, P. G. (1986, November). _Some conditions affecting assessment of job requirements_. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Smith, E. P., & Walker, C. B. (1987, November). _Short versus long term tenure as a criterion for validating biodata_. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.)

Wing, H., Hough, L. M., & Peterson, N. G. (1987, August). <u>Predictive validity of noncognitive measures for Army classification and attrition</u>. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.)

Wise, L. L., Campbell, J. P., & Peterson, N. G. (1987, April). <u>Identifying optimal predictor composites and testing for generalizability across jobs and performance constructs</u>. Paper presented at the Annual Conference of the Society of Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.) <u>Personnel Psychology</u>, 43, 2, Summer 1990

Wise, L. L., McHenry, J. J., Rossmeissl, P. G., & Oppler, S. H. (1986, November). <u>ASVAB validities using improved job performance measures</u>. Paper presented at the Annual Conference of the Military Testing Association, Mystic, CT. (In ARI Research Note 88-23.) <u>Personnel Psychology</u>, 43, 2, Summer 1990.

Wise, L. L., McHenry, J. J., & Young, W. Y. (1986, December). <u>Project A Concurrent Validation: Treatment of missing data</u>. (ARI Selection and Classification Technical Area Working Paper 86-08.)

Young, W. Y., Harris, J. H., Hoffman, R. G., Houston, J. S., & Wise, L. L. (1987, April). <u>Large scale data collection and data base preparation</u>. Paper presented at the Conference of the Society of Industrial and Organizational Psychology, Atlanta. (In ARI Research Note 88-23.) <u>Personnel Psychology</u>, 43, 2, Summer 1990.

Fiscal Year 1988

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army officer sample. Organizational Behavior and Human Decision Process, 40, 307-322.

Campbell, J. P. (Ed.). (1987, October). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1987 fiscal year (HumRRO IR-PRD-88-18). (ARI Technical Report 862.)

Campbell, J. P., & Harris, J. H. (1988, August). The Project A approach to describing job proficiency. Paper presented at the Annual Convention of the American Psychological Association, Atlanta.

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute. (1988). Improving the selection and classification of Army enlisted personnel: Annual Report, 1987 Fiscal Year - Supplement to ARI Technical Report 862. (ARI Research Note 88-23).

McHenry, J. J., & Felker, D. B. (1988, August). Assessment of problem-solving skills via traditional hands-on and knowledge tests. Paper presented at the Annual Convention of the American Psychological Association, Atlanta.

Nord, R. D., & White, L.A. (1988). The measurement and application of performance utility. In B. Green, H. Wing, & A. Wigdor (Eds.), Linking Military Enlistment Standards to Job Performance, pp. 215-243. Washington, DC: National Academy Press.

Sadacca, R., Campbell, J. P., White, L. A., & DiFazio, A. S. (1988, July). Weighting criterion components to develop composite measures of job performance (HumRRO IR-PRD-88-14). (ARI Technical Report 838.)

Sadacca, R., White, L. A., Campbell, J. P., DiFazio, A. S., & Schultz, S. R. (1988, August). Assessing the utility of MOS performance levels in Army enlisted occupations (HumRRO IR-PRD-88-15). (ARI Technical Report 83S.)

Fiscal Year 1989

Campbell, J. P. (Ed.). (1989, December). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report 1988 fiscal year (HumRRO IR-PRD-89-28). (ARI Technical Report in preparation.)

Campbell, J. P., & Zook, Lola M. (Eds.). (1990). Improving the selection, classification, and utilization of Army enlisted personnel: Final Report on Project A (HumRRO FR-PRD-90-06). (ARI Technical Report in preparation.)

Hansen, M. A., & Borman, W. C. (1989, April). Development and construct validation of a situational judgment test as a job performance measure for first line supervisors. Paper presented at the Annual Convention of the Society of Industrial and Organizational Psychology, Boston.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J., & McCloy, R. (In Press). Criterion-related validities of personality constructs and the effect of response distortion on those validities. Journal of Applied Psychology.

Nord, R. D., & White, L. A. (1989). Performance utility and optimal job assignment (ARI Technical Report in preparation).

Olson, D. M., & Borman, W. C. (1989). More evidence on relationships between the work environment and job performance. Human Performance, 2(2), 113-130.

Pulakos, E. D., Borman, W. C., & Hough, L. M. (1988). Test validation for scientific understanding: Two demonstrations of an approach to studying predictor-criterion linkages. Personnel Psychology, 41, 703-716.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). An examination of race and sex effects on performance ratings. Journal of Applied Psychology, 74, 770-780.

Appendix C

PROJECT A EMPLOYEES

Project A Employees

Paula Abato
Lillian Adderson
Dave Albright
Brad Allen
Jane Arabian
Bob Archer
Louis Armijo
Terry Armstrong
Steve Ashworth
Lance Atiyeh
Albert Atkins
Kathy Austin

Katy Bair
Tom Bakke
Elaine Balek
Dennis Baltzley
Barbara Barbosa
Bruce Barge
Jeff Barnes
Fred Barnette
Bill Baylor
James Beckham
Tracy Benton
Cindy Beresh
Theresa Berry
Philip Bobko
Wally Borman
Mike Bosshardt
Joyce Bowers
Ted Bowler
Richard Bradfield
Allen Bremwell
David Brandt
Tim Bronson
Kimberly Brooks
Diane Brown
Melinda Browne
Suzanne Browne
Sandra Buenahora

Bobbie Cadwell
Chareline Cain
Charlotte Campbell
John Campbell
Kimberly Campbell
Roy Campbell
Dave Campshure
Cay Carn
Michael Carrigan
Gary Carter
Al Castelli
Jana Challberg
Wei Jing Chia
John Claudy

Grady Coats
Steve Coats
Adrienne Colella
Ronald Coltress
John Connors
Tom Cook
Vy Vy Corpe
Henry Counts
Jennifer Crafts
Pam Croom
Linda Culp
Denita Curry
Mark Czarnolewski

Arnold Daniels
J. L. Darrah
Greg Davis
Gene Davis
Ralph Davis
Bob Davis
Rosemary Dawson
Robin Dean
Tommy deGrom
Katie Delaplane
Donna Denard
Vernell Denman
Natalie Depp
Darren Dickens
Ani DiFazio
Denise Dinnen
Gwendolyn Dixon
Theresa Doty
Kim Downing
Jack Doyle
Elizabeth Doyle
Eugene Drucker
DuBois, Dave
Mary Duffy
Mike Dunn
William Dunn
Marv Dunnette

Kent Eaton
Jo Edwards
Ed Eisner
Tom Eissenberg
Leon Elder
Ernesto Endaya
Benita Evaro

Chris Felker
Dan Felker
Lacy Ferguson
Stephanie Fields
Santiago Fierro

Janice Fisher
Sub Fisicaro
Brent Fleming
Pat Ford
Max Foster

Barnett Gambel
Ilene Gast
Nancy Gazzerro
Mike Geyer
Kevin Gilmartin
Marv Goer
Steve Goffard
Santos Gonzalez
Charlene Gower
Francis Grafton
Francis Gragg
Bernie Green
Miranda Green
Kay Grimes
Monica Gustus
Gloria Guth

Joe Hackney
John Hagans
Cliff Hahn
Milton Hakel
Glenn Hallam
Barbara Hamitlon
Laurel Hamilton
David Hannaman
Larry Hanser
Mary Ann Hanson
Chuck Harnest
Andrea Harris
Carolyn Harris
Delaine Harris
Jim Harris
Ingrid Heinsohn
Susan Heon
Brian Hilburn
Carolyn Hill
Susan Hill
Michael Hillelsohn
Rose Hiles
Candace Hoffman
Gene Hoffman
Peggy Hoffman
Anna Holcomb
Sandy Holland
Kathy Holloway
Berkeley Holmes
William Holmes
Suzanna Horvath
Leaetta Hough

Marlou Hough
Merv Hough
Janis Houston
Nancy Huffman
Harold Hull
Lloyd Humphreys
Tom Hydock

Robert Jagers
Emma James
Dick Jamieson
Gregory Jefferson
Lawrence Johnson
Pam Johnson
Sallie Johnson
Scott Johnson
Edna Johnston
Agnes Jones
Denise Jones
Gail Jones
John Joyner

John Kamp
Ken Keiser
Sue Keskinen
Meg Keyes
Mary Kirkman
Deirdre Knapp
Tom Kracker
Manuela Krass
Bob Krug
Dick Krulik
Ruthene Krulik
Doug Kuhn
Susan Kushner

Steve Lammlein
Chris Larsen
Alan Lau
Pat Lawler
Debra Lewis
Kathy Lillie
Colleen Lincoln
Timothy Ling
Robert Linn
Carmen London
Jim Lucas

Wynn MacDonald
Rod McCloy
Marshall McClintock
Tim McCollum
Renee McCord
Deborah McDaniels
Dan McGarvey

Allison McGrady
Matt McGue
Lesley McHale
Jeff McHenry
Don McLaughlin
Brian McNelis
Carol Manning
Debbie Marcum
Clessen Martin
Derrick Martin
Fran Martin
Mary Martin
Scott Martin
Gary Maus
Betty May
Mary Medved
Vincent Melillo
Jane Mell
Ray Mendel
Dolores Miller
Yvonne Miller
Jeanetta Milliner
Sue Milnor
Carol Miner
Karen Mitchell
Elizabeth Moore
Joe Moore
Rita Morley
Steve Motowidlo
Rebecca Mordini
Steve Morris
Greg Mosher
Debbie Myers

Sandy Napier
Brett Newson
Dianne Nilsen
John Novak

Bridgette O'Brien
Leonardo Olis
Joe Olmstead
Darlene Olson
Scott Oppler
Bill Osborn
Milquella Otero
Cyndi Owens-Kurtz
Mark Owens-Kurtz
Rebecca Oxford-
   Carpenter

Bruce Palmer
Pat Parham
Mavee Park
Cheryl Paullin
Bridget Peoples
Anthony Perez

Norm Peterson
Scott Peterson
Ilsa Pineda
Beverly Popelka
Kent Porter
Ronald Poulson
Joann Proctor
Phyllis Pryse
Donovan Puffer
Elaine Pulakos

Paul Radtke
David Reis
Jodi Reynolds
Barry Riegelhaupt
Robin Riegelhaupt
Diana Kay Ring
Dave Rivkin
Lori Roberson
Michael Robinson
Don Rogan
Andy Rose
Sharon Rose
Harvey Rosenbaum
Richard Rosenblatt
Rod Rosse
Paul Rossmeissl
Michael Rumsey
Darlene Russ-Eft
Teresa Russell

Bob Sadacca
Lelia Sanders
Martha Sanders
Susan Schechtman
Liddy Schneider
Sheila Schultz
Aurelia Scott
Cindy Seale
Jeannette Sekellick
Batia Sharon
Betty Shelley
Donna Shepherd
Harris Shettel
Joyce Shields
Paula Singleton
Nancy Skilling
Deb Skophammer
Elizabeth Smith
Helen Sperling
Bill Spruill
Jim Stallings
Louis Stamps
Cathy Stawarski
Brian Stern
Susie Stern
Gayle Stifle

Ron Still
Arthur Stone
Melanie Styles
Rod Symmes
Phil Szenas

Zarva Taru
Connie Taylor
Elaine Taylor
Mary Anne Taylor
Mary Tenopyr
Barb Thomas
Alice Thompson
Tom Tiffany
Doris Torres
Jody Toquam
Dick Trotter
Suzanne Tsacoumis
Carmen Tyminski

Jay Uhlaner
Richard Urich

Theresa Van Noy
Robert Vineberg

Andrew Walton
Ming-mei Wang
Ray Weinberg
Linda Wells
George Wheaton
Debbie Whetzel
Clint Walker
Leon Wetrogan
Len White
Todd Will
Anne Williams
Walter Williams
Effie Wilson
Thurlow Wilson
Hilda Wing
Laurie Wise
Ronita Wisniewski
Helen Woodard
Diane Woodrum
Marcia Wojsko
Becky Wyant
Carol Wyman

Winnie Young

Laurie Zaugg
Ray Zimmerman
Lola Zook